UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Implementation and evaluation of ratio-based conditional average treatment effects in high-dimensional omics.

Anthony Okumu

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :
Prof. dr. Olivier THAS

SUPERVISOR :
Koen VAN DEN BERGE
Oliver DUKES

2023
2024

Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Implementation and evaluation of ratio-based conditional average treatment effects in high-dimensional omics.

Anthony Okumu

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Bioinformatics

SUPERVISOR :

Prof. dr. Olivier THAS

SUPERVISOR :

Koen VAN DEN BERGE

Oliver DUKES

Acknowledgements

My heartfelt thanks to my supervisors, Dr. Koen Van den Berge, Prof. Dr. Oliver Dukes, and Prof. Dr. Olivier Thas, for their invaluable guidance and support in my thesis journey.

In a special way, I would also like to express my appreciation to VLIR-UOS for their financial support, which has enabled me to pursue a two-year master's program at the University of Hasselt in Belgium.

My Joy knows no bounds in expressing my warm appreciation to my beloved parents Mr. and Mrs. Okumu John. You have been my number one source of encouragement.

Above all, great honor and glory be to God the Almighty for His knowledge, wisdom, and blessings to successfully have made this possible.

Abstract

The estimation of exposure effects in observational studies is often complicated due to confounding factors, particularly in high-dimensional 'omics data. The traditional regression framework of adjusting for potential confounders by simply adding them along with the exposure to the outcome model may not adequately address the issue of confounding.

In this study, we perform causal differential gene expression on single-cell RNA-sequencing data comparing lupus patients versus healthy controls. The study focuses on estimating ratio-based causal disease effects on gene expression, specifically exploring different approaches in estimation of causal effects. Before estimation of causal effects, the impact of confounding factors of disease effects on gene expression is explored.

Inverse probability weighting, standardization and doubly robust estimators are explored to estimate the average causal effect as a ratio. Furthermore, this project details the implementation and evaluation of a novel ratio-based approach; the contrast regression, for estimating conditional average treatment effects in a genomics setting.

From the results, confounding factors of disease effect on gene expression; Age, Sex, Ancestry, and Batch have an impact on the effect size estimates and the number of significant genes. In addition to this, it is also interesting to note that the impact of confounding on the effect size estimates varies from one gene to another. Our findings suggest that the choice of the estimator is very important, as evidenced by totally different estimates for the same estimand based on different estimators while considering the same gene. The findings further suggest that inverse probability weighting yields closer estimates to the doubly robust method as compared to standardization. Additionally, the contrast regression ratio-based approach yields individualized conditional average treatment effects scores from which effect heterogeneity can be explored among individuals. The approach offers avenues for integration of machine learning methods which are well-suited for analyzing high-dimensional genomic data. However, incorporating these machine learning algorithms may require more computational time.

key Words: **Causal inference, Confounding, Inverse Probability Weighting, Standardization, Doubly Robust Estimator, Contrast regression, Two regression**.

Contents

1 Introduction

1.1 Background

In many medical-related studies, particularly drug evaluations or understanding disease effects, it is critical to estimate the effects of treatments on outcomes by contrasting comparable patient groups. When the groups being compared (such as the diseased group and the healthy group) differ systematically, confounding factors can cloud the understanding of the effects of the treatment or disease. For instance, if the treatment group is generally older, predominantly consists of one gender group while the healthy group is mostly the opposite gender, or primarily composed of individuals from one ethnic background and the healthy group from another. Randomized Clinical Trials (RCTs) address these issues by design, as they randomly assign participants to various treatment groups to prevent such discrepancies. The random assignment of participants to treatment and control groups in an RCT ensures that the estimated treatment effect is unbiased, as it is not influenced by confounding factors (Lee and Lee, 2022). Comparing outcomes across randomized treatment arms yields an intention-to-treat effect, a robust measure of causal effect (Goetghebeur et al., 2020).

While RCTs are considered the gold standard, they are not always feasible or ethical for every research question (Smith et al., 2022). For example, consider a study designed to evaluate treatments for a hereditary heart condition. It would be unethical to randomly assign patients to either heart surgery or non-surgical treatments like medication. Instead, doctors tailor treatments based on the patient's specific health needs and the latest scientific evidence. Consequently, observational studies play a critical role in understanding treatment effects in real-world settings when randomization is not feasible.

In genomic studies, researchers frequently aim to investigate the effect of an exposure on genomic variables like gene expression (Reifeis et al., 2020). Due to ethical reasons, conducting RCTs on human populations is not feasible in these scenarios. Thus, observational studies become a valuable alternative in estimating the exposure effects on an outcome.
However, these studies introduce the challenge of confounding, where differences between the compared exposure groups can obscure the true causal effect of the exposure (Imbens and Rubin, 2015). This means that they typically provide estimates of associations rather than direct causal

links (Altman and Krzywinski, 2015). Traditional regression framework of adjusting for potential confounders by simply adding them along with the exposure in a regression model may not adequately address the issue of confounding variables and does not directly produce the estimate of interest (Reifeis et al., 2020).

Causal inference is concerned with analysis that allows one to interpret the estimated effects as causal (effects directly caused by the exposure). The "causal inference" framework introduced by Neyman (1923) and Rubin (1974) formally defines causal effects in terms of *counterfactuals*, or *potential outcomes*, which extends the causal inference framework from randomized experiments to observational studies. In this framework, one could contrast the expected outcome variable if everyone had received treatment versus if everyone had received placebo. Suppose a study population where an outcome variable of interest is denoted by Y and a dichotomous treatment D (with $D = 1$ if an individual is treated and $D = 0$ if an individual is not treated). Let $Y^{(1)}$ be the outcome that would have been observed under treatment ($D = 1$) and $Y^{(0)}$ be the outcome that would have been observed under no treatment ($D = 0$). $Y^{(0)}$ and $Y^{(1)}$ are referred to as potential outcomes (sometimes referred to as counterfactual outcomes). For an individual i, either of the counterfactual outcomes can be observed ($Y_i^{(0)}$ or $Y_i^{(1)}$); the one corresponding to the actual treatment received by the individual. This creates a causal missing data problem since we care about making causal claims where we contrast between the potential outcomes. Due to the fact that individual causal effects cannot be identified, the focus turns to average causal effect.

In order to answer a specific research question, one needs to predefine the target parameter (an estimand); the unknown quantity of interest. Commonly, causal estimands are expressed in several forms: as differences (on the additive scale) or as ratios (on the multiplicative scale) to quantify the causal effect of interest. Table 1 shows the risk difference and relative ratio estimands in terms of expected potential outcomes.

Table 1: Causal estimands for risk difference and relative ratio contrasts.

Effect Measure	**Marginal (ATE)**	**Conditional (CATE)**
Risk difference	$\mathrm{E}\left[Y^{(1)} - Y^{(0)}\right]$	$\mathrm{E}\left[Y^{(1)} - Y^{(0)} \mid Z = z\right]$
Relative ratio	$\frac{\mathrm{E}[Y^{(1)}]}{\mathrm{E}[Y^{(0)}]}$	$\frac{\mathrm{E}[Y^{(1)} \mid Z=z]}{\mathrm{E}[Y^{(0)} \mid Z=z]}$

Note: **ATE**: Average Treatment Effect; **CATE**: Conditional Average Treatment Effect; Z: Covariates.

Conditional Average Treatment Effect (CATE) is a preferred estimand when effect heterogeneity within subgroups defined by covariates Z is of interest. Several methods (estimators) have been developed for estimating causal effect measures. It is important to note that we rely on certain strong assumptions in order to infer causation from the observed data (Rosenbaum and Rubin, 1983).

Addressing the issue of confounding in observational genomics studies is paramount while investigating the true causal effect of the exposures on gene expression, although this has received relatively less attention. Studies such as Reifeis et al. (2020) and Hejazi et al. (2023) focused on assessing exposure effects in a genomics setting by estimating the average causal effects focusing on continuous outcomes. Recent work by Du et al. (2024) focuses on semiparametric causal inference approaches applicable to Single-cell CRISPR data. Additionally, several studies not necessarily focussing on genomics; [Xie et al., 2012; Wager and Athey, 2018; Künzel et al., 2019, and others] have explored estimating the CATE for continuous outcomes. All these studies have focused on assessing the absolute difference (on an additive scale) between potential outcomes in the contrast groups, that is, $\mathrm{E}\left[Y^{(1)} - Y^{(0)} \mid Z = z\right]$ to quantify causal effects. However, with most data from epidemiological studies and high-throughput sequencing technologies, where the outcome (such as gene expression levels) are typically counts, focusing on differences may not be suitable.

In studies where one encounters count or binary outcomes, effect measures on a multiplicative scale, i.e., ratios, are more appropriate. A few studies, such as Dukes and Vansteelandt (2018) work on G-estimators, highlight on the estimation of multiplicative causal contrasts (as risk ratios for binary outcomes). Therefore, investigating ratio-based treatment effects for count outcomes

merits further exploration. In response to this need, Yadlowsky et al. (2021) proposed a statistical methodology for the estimation and validation of the ratio-based CATE, $\frac{\mathrm{E}[Y^{(1)}|Z=z]}{\mathrm{E}[Y^{(0)}|Z=z]}$ suitable for count outcomes on a set of confounders Z. This method, building on the semiparametric causal inference literature, notably the work of Robins and Rotnitzky (2001), Van der Laan and Rose (2011), and Dukes and Vansteelandt (2018), integrates flexible machine learning methods well-suited for high-dimensional datasets.

1.2 The Study Objectives

This study focuses on the estimation of ratio-based CATE on a single-cell gene expression lupus dataset published in Perez et al. (2022), comparing lupus patients to healthy controls. Specifically our study implements the proposal by Yadlowsky et al. (2021) and assesses its computational efficiency in estimating ratio-based CATE. We evaluate the impact of confounding factors of disease effects on gene expression. Further, we contrast results obtained based on the different approaches on estimating the disease effect within the causal inference framework.

The experimental design, data description as well as statistical methodologies used are discussed in Section 2. The results from the analyses are presented in Section 3. Discussion on the results follows in Section 4. Section 5 provides the ethical thinking, societal relevance, and stakeholder awareness of the study. Finally, conclusions and recommendations for future research are made in Section 6.

2 Methodology

2.1 Data

2.1.1 Single-cell RNA-sequencing

Single-cell RNA-sequencing (scRNA-seq) is a widely used and powerful technology that enables profiling of the transcriptomes of numerous individual cells (Andrews et al., 2021). It quantifies mRNA molecules in individual cells. In general, the workflow begins with extracting tissue samples, which are broken down during single-cell dissociation. Cells are then isolated individually, either into wells using plate-based methods or captured within microfluidic droplets (Mereu et al., 2020). The resulting mRNA sequences are aligned to a reference genome using cellular barcodes or unique molecular identifiers (UMIs) to obtain a count matrix that details gene expression by cell (Heumos et al., 2023). The generated expression matrix representing the number of transcripts observed for each gene and cell is considered the starting point for our analysis workflow. Identifying systematic transcriptional changes across conditions through differential expression analysis is a key aspect of analyzing scRNA-seq data (Stegle et al., 2015). These analyses are essential for gaining insight into the molecular responses that occur during development, following perturbations, or in various disease states (Kotliar et al., 2019).

2.1.2 Data Description

Systemic lupus erythematosus (SLE), commonly referred to as lupus, is a chronic autoimmune disease that can affect various body organs such as the skin, showing a higher incidence in women and populations of Asian and European ancestries (Carter et al., 2016). In our study, we analyze a single cell gene expression - lupus data set described and analyzed in Perez et al. (2022). It is an extensive collection of scRNA-seq data from Peripheral Blood Mononuclear Cells (PBMCs). The dataset includes over 1.2 million PBMCs from 261 unique samples (162 SLE cases and 99 healthy controls), covering individuals of Asian, European, African American and Hispanic ancenstry. The data was generated using multiplexed scRNA-seq (mux-seq), a technique developed to enable systematic, cost-effective profiling of large population cohorts with reduced variability. The 261 samples and 91 replicates were profiled in 23 pools across 4 processing cohorts. Following quality control and removal of doublets and other contaminants such as platelets and red blood cells, the final dataset was refined to contain 1,263,676 cells. The data is

stored as a SingleCellExperiment object containing the single-cell data - gene-by-cell expression data, per-cell metadata, and per-gene annotation. The cell metadata contains variables including; `Batch (batch_cov)`, `individual id (ind_cov)`, `cell type (cg_cov)`, `Age`, `Sex`, `Ancestry (pop_cov)`, and `SLE status`. The single cells are clustered into 11 biological cell types.

For our analysis, we consider a subset of the data for only a single cell type (the classical monocytes or cM), which consists of 307,429 cells with a selected set of 12,869 genes that are expressed in at least 300 cM cells. Further, we ignored samples from African American and Hispanic since they consist of few data points (3 from the African American and 2 from Hispanic ancestry). Only samples from Asian and European ancestry were considered.

2.2 Data Processing and Preparation

Most scRNA-seq differential expression methods are currently designed to compare cell subpopulations or conditions, treating individual cells as experimental units for statistical modeling. However, with the proliferation of multi-sample, multi-group scRNA-seq datasets involving many cells per sample, the focus shifts towards making sample-level inferences by considering samples as the experimental units (Crowell et al., 2020). Pseudo-bulk samples are created by aggregating read counts together for all the cells in each cell type and by sample. The strategy of aggregating cell-level data into "pseudobulk" counts and employing statistical models for sample-level inferences enhances the capacity to interpret complex scRNA-seq datasets. Simulation studies by Crowell et al. (2020) have hdemonstrated that the aggregation-based differential expression methods are both highly computationally efficient and stable, thereby outperforming those designed specifically for single-cell level. Based on their study, it was concluded that although Mixed Model methods perform comparably well, their high computational burden may not justify the flexibility they offer. Further, by aggregating cells within each replicate, pseudobulk methods dramatically reduce the number of zeros in the data, especially for lowly expressed genes (Squair et al., 2021).

2.2.1 Creating Pseudobulk Samples

We aggregated the single cell level counts to sample level by summing counts together for all cells with the same combination of batch and sample IDs to generate pseudo-bulk expression counts

for every gene. Batch variable was considered while aggregating samples due to the fact that our dataset contains multiple samples from the same individual, and therefore, it could happen that some samples for the same individual belong to different batches. Thus, simply adding counts of the same sample would ignore the batch effect. The differential gene expression analysis was thus performed based on the "pseudobulk" samples.

2.2.2 Independence Assumption

In our dataset, multiple samples were taken from the same individual. However, these samples are inherently correlated because they come from the same individual, violating the independence assumption central to Generalized Linear Models (GLMs). GLMs presuppose that each observation (in this case, each sample) is independent of the others. We thus considered selecting a single sample per patient for analysis in order to simplify the dataset to a form where the independence assumption of GLM is adhered to. In our case, we choose to retain samples with most amount of information. That is, for multiple samples from the same individual, we consider the sample with the highest total count since we are working with a single-cell RNA-sequencing protocol adopting unique molecular identifiers (UMIs).

2.2.3 Gene Filtering

Gene filtering is a potent approach involving the removal of genes unlikely to exhibit differential expression. This process significantly boosts the sensitivity of analyses and computational efficiency, yielding more meaningful results (Gohlmann and Talloen, 2009). Genes that are lowly expressed can be detected based on the Count Per Million (CPM) cutoff defined as;

$$\text{CPM.cutoff} = \frac{\text{min.count}}{\text{median(lib.size)} \times 10^6} \ .$$

Genes are kept if their CPM $\geq$ CPM.cutoff computed using the `filterByExpr()` edgeR function. In addition, each kept gene is required to have the minimum count of 15 reads (the default option) across all the samples.

2.2.4 Normalization

When analyzing differences in gene expression across samples, correcting for the differences in library sizes is essential in order to avoid composition biases between libraries. Library size here

refers to the total number of read counts for a given sample. If, in a sample, there are a few genes that are highly expressed and take up a substantial proportion of the total library size, the remaining genes will be undersampled. This means that the remaining genes may incorrectly appear downregulated in that sample (Robinson and Oshlack, 2010). Normalization methods can thus be used to adjust for such differences.

Prior to differential gene expression (DGE) analysis, normalizing factors were computed based on Trimmed Mean of M-values (TMM). TMM normalization involves determining a set of genes (that are not DE) by trimming 30% of M-values($\log_2$ foldchange between samples) and 5% A-values(Average log normalized counts). This subset of genes are then used to calculate the normalizing factors for each sample by taking the weighted mean. Normalization factors for each sample based on TMM were calculated using the `calcNormFactors()` function from the `edgeR` package. The computed factors were then used to obtain the normalized library sizes, i.e., normalized library sizes = original library sizes * normalization factors. The (log-transformed) normalized library sizes are then brought into the GLM model as an offset.

2.3 Statistical Analysis

2.3.1 Modelling Gene Expression Counts

The resulting expression matrix of aggregated sample level data consists of counts observed for each gene and sample. Often, count data is modeled using a Poisson distribution. This distribution relies on the special property that the variance is equal to the mean, that is,

$$\lambda = \mathrm{E}(Y) = \mathrm{Var}(Y).$$

However, the variance of gene expression across multiple biological replicates often exceeds the mean expression values, indicating overdispersion. In order to account for overdispersion, a negative binomial distribution is often used, which incorporates an overdispersion parameter, ϕ, that controls the amount of overdispersion in the data. Therefore, we assume that

$$Y_{gi} \overset{i.i.d}{\sim} NB(\lambda_{gi}, \phi_g),$$

with

$$\mathrm{E}(Y_{gi}) = \lambda_{gi}, \quad \mathrm{Var}(Y_{gi}) = \lambda_{gi} + \phi_g \lambda_{gi}^2.$$

where, Y_{gi} is the expression levels of a gene g for sample i, λ_{gi} is the mean count and ϕ_g - the overdispersion parameter of a gene g. A Generalized Linear Model (GLM) with a negative binomial distribution assumption is thus used to model the expected counts via a log-link function.

One can conveniently perform DGE analyses on pseudo-bulk samples using the `pseudoBulkDGE` function in `edgeR` to model the counts directly. However, for our subsequent analyses based on the study objectives (to incorporate causal inference methods), we typically need weights in the GLM model, and weights in `edgeR` have a different interpretation. Therefore, we opt to fit gene-by-gene Negative Binomial (NB) models using the `glmmTMB` package, which offers fast and stable fitting of the NB-GLM models (Brooks et al., 2017).

2.4 Causal Inference

Causal inference is concerned with analysis that allows one to interpret the estimated effects as causal under certain restrictions and assumptions. The data set that we are analyzing is typically from an observational study where Age, Sex, and Ancestry can confound the disease effect on gene expression. In this case, we cannot therefore make causal claims (that effects are directly caused by the disease) from the traditional regression methods, but rather associations. Figure 1 illustrates the distinction between causation and association where the population under study (represented by a diamond) is divided into the diseased and non diseased (healthy) groups.

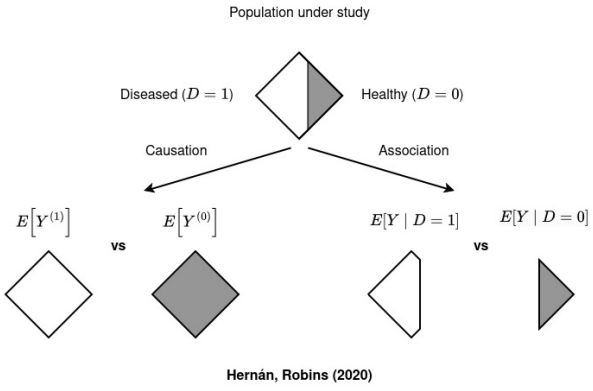


Figure 1: An illustration showing the difference between causation and association. The population represented by a diamond is segmented into a larger white section (the diseased group) and a smaller grey section (the healthy group).

Figure 1 depicts that, causation involves comparing the same population under the alternative exposure scenarios whereas association involves comparing two separate subsets of the population. One major reason as to why we are hesitant to attribute causal interpretations to observational associations stems from the absence of randomized treatment assignment (Hernan and Robins, 2020). Unlike randomized studies, where participants are assigned to treatment or control groups through a random process to ensure comparability and minimize bias, observational studies typically do not involve such controlled assignments. This lack of randomization means that the groups being compared may differ in ways other than the exposure of interest, potentially confounding the results and making it difficult to ascertain causality purely from observational data. We will hereafter denote disease status (i.e. diseased or nondiseased) as the binary exposure variable D with levels 1 and 0, gene expression as the outcome variable Y, and a set of confounding variables Age, Sex, batch and Ancestry all denoted as Z. Figure 2 depicts a directed acyclic graph showing the causal structure in which a set of factors Z confound the direct effect of exposure D on the outcome Y. Z induces (non-causal) association between D and Y.

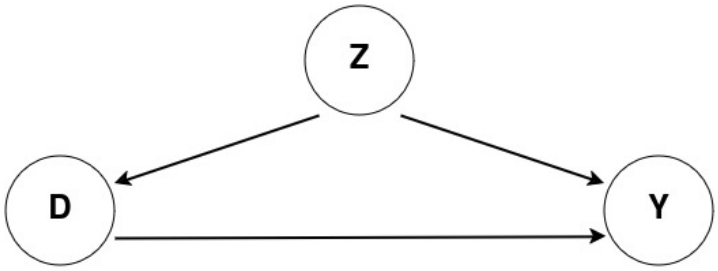


Figure 2: A directed acyclic graph illustrating the causal structure. Y: outcome variable; D: a binary exposure variable; Z: a set of variables to control for confounding.

Based on the potential outcomes framework, introduced by Neyman (1923) and Rubin (1974), which extends causal inference framework from randomized experiments to observational studies, causal effects can be defined in terms of the potential outcomes $Y^{(1)}$ and $Y^{(0)}$; where
$Y^{(1)}$ denotes the outcome variable that would be observed under $D = 1$ and $Y^{(0)}$ denotes the outcome variable that would be observed under $D = 0$.

In order to obtain causal effects, one must make contrast between the potential outcomes in the given exposure groups. In reality, only one of the poqzderdrtential outcomes can be observed

for an individual. For instance, one cannot be both diseased and healthy at the same time. This creates a missing data problem which obstructs the identification of the individual causal effects since they cannot be expressed as a function of the observed data (Hernan and Robins, 2020). However, relying on certain assumptions, we can identify potential outcomes from the observed data to estimate the average causal effect (also known as Average Treatment Effect, ATE) (Robins, 1986). We therefore make the following assumptions (as described in Hernan and Robins (2020)) in order to estimate the average causal effect based on the observed data.

2.4.1 Assumptions

Consider the binary exposure variable D , observed outcome variable Y and a set of confounders Z defined as earlier in the text;

① **Consistency:** The assumption of consistency relates the observed outcomes to the potential outcomes. Consistency holds if the observed outcome for all individuals is the same as the potential outcome that would be realized in response to setting the treatment to the exposure level that was observed. It can be formally written as;

$$Y = DY^{(1)} + (1 - D)Y^{(0)}.$$

We further assume that observations are independent (no interference) and there is no measurement error.

② **Conditional Exchangeability:** Randomized studies ensure that conditional and marginal exchangeability hold by design since the independent predictors of the outcome are equally distributed between the exposure groups. However, when treatment is not randomly assigned, the reasons for receiving treatment are likely to be associated with some covariates. Consequently, conditional exchangeability cannot be guaranteed in observational studies but can be assumed to hold if the unmeasured risk factors affecting outcomes are evenly distributed among exposure groups, given the measured confounders are controlled for. Thus, conditional exchangeability can be formally stated as;

$$\left\{Y^{(1)}, Y^{(0)}\right\} \coprod D|Z.$$

In otherwords, there is no confounding within levels of Z since controlling for Z has made the exposure groups comparable. This is referred to as the unconfoundedness assumption

(Imbens and Rubin, 2015). We assume in our analysis that Z is a sufficient set of variables that confound the relationship between disease and gene expression.

③ **Positivity:** Positivity holds if the conditional probability of being in the exposure groups is greater than zero; i.e.,

$$\text{If } Pr(Z = z) > 0, \text{ then } 0 < Pr(D = d \mid Z = z) < 1; \forall z \in Z, d \in \{0, 1\}.$$

Our study aims at performing causal differential gene expression therefore seeking to contrast between the expected outcome if all individuals had the disease versus if all individuals had no disease based on each gene. With the assumptions stated above, the observed data can therefore be used to estimate the average causal effect. For count data, effect measures on a multiplicative scale i.e, ratios are more appropriate such as the marginal ratio;

$$B = \frac{\mathrm{E}[Y^{(1)}]}{\mathrm{E}[Y^{(0)}]}, \tag{1}$$

or as a conditional ratio;

$$B(z) = \frac{\mathrm{E}[Y^{(1)} \mid Z = z]}{\mathrm{E}[Y^{(0)} \mid Z = z]}. \tag{2}$$

The estimands in Equations 1 and 2 contrast the expected potential outcomes in exposure groups $D = 1$ versus $D = 0$. Equation 1 is referred to as the Average Treatment Effect (ATE) and Equation 2 is referred to as the Conditional Average Treatment Effect (CATE).

2.4.2 Causal Effect Estimation Approaches

Inverse Probability Weighting (IPW)

In observational studies, exposure is often more likely for certain individuals than for others (Smith et al., 2022). For instance suppose that in our study, younger individuals have an unusually high incidence of being diseased. IPW is one of the standard approaches that can be used to balance the differences in the distribution of the measured covariates Z in the exposure groups being compared. In this context, we compute the weights of each individual by the inverse of their probability of exposure, referred to as the propensity score. The weighting is such that individuals in the treated group who had a low chance of being treated are up-weighted, and likewise, it upweights those in the untreated who had a low chance of being untreated (Horvitz and Thompson, 1952). This creates a "pseudo-population" in which the distribution of the

measured covariates, Z, is identical in both treatment groups (Hernán and Robins, 2020) and thus comparable. For instance, the diseased and healthy have the same age distribution on average after weighting. IPW estimation involves the following steps;

① **Fitting the propensity score model (i.e, a logistic regression model for a binary exposure)**.

$$\text{logit}(Pr(D = 1|Z)) = \gamma_0 + \gamma_1^T Z, \tag{3}$$

where γ_0 is the intercept and γ_1 is vector of parameter coefficients corresponding to covariates $z \in Z$.

Use the fitted model to obtain the weights, which are the inverse of the predicted probabilities of receiving the treatment that an individual actually received. Stabilized weights are often preferred since they yield narrower confidence intervals (Hernán and Robins, 2020). These can be calculated as

$$\hat{W}_i(d) = d\ \frac{\hat{Pr}(D_i = 1)}{\hat{Pr}(D_i = 1|Z)} + (1 - d)\ \frac{\hat{Pr}(D_i = 0)}{\hat{Pr}(D_i = 0|Z)},\ d = \{0, 1\}. \tag{4}$$

② Fitting the weighted regression outcome model given treatment to estimate the causal effect.

Under the standard assumptions of causal inference; conditional exchangeability, positivity and consistency, Association is causation in the pseudo-population. We further assume that the treatment model in Equation 3 is correctly specified.

Standardization (g-formula)

Standardization is an alternative approach to IPW, where data is first expanded into three copies (Hernán and Robins, 2020). The first copy is the original dataset. In the second and the third copies, all individuals are included but with their exposure groups D set to 0 (un exposed) and 1 (exposed) respectively and the outcome data is deleted. Using the first copy of the data, we obtain the parameter estimates by fitting a regression model for the outcome, $E\left[Y|D = d, Z = z\right]$ in each of the exposure groups given the confounders Z. In the two other copies, the parameter estimates obtained from the first copy are then used to generate the predicted outcome in the exposed ($D = 1$) and in the unexposed ($D = 0$) considering the same covariates as in the original data. This yields, the predicted means in each of the exposure groups, i.e.

$$\hat{E}\left[Y|D = 1, Z = z\right] \text{ and } \hat{E}\left[Y|D = 0, Z = z\right].$$

Note that every individual, i has two predicted values corresponding to the two estimated potential outcomes. We thus, compute the average as $\frac{1}{n}\ \sum_{i=1}^{n} \hat{E}\left[Y_i|D_i = d_i, Z_i = z_i\right], d_i \in \{0,1\}$ in each of the exposure groups over all individuals and estimate the average causal effect in Equation 1 by contrasting the estimated average potential outcomes in the exposure groups.

In summary, standardization involves three key steps: Model fitting, predicting, and averaging. Like IPW, the standardization approach relies on the standard assumptions of causal inference mentioned earlier in the text. Under these assumptions, the standardized mean outcome in the (un)exposed is a consistent estimator of the mean outcome if everyone had been (un)treated (Hernan and Robins, 2020). We also assume that the outcome model is correctly specified.

Doubly Robust Methods

It has been discussed above that both IPW and standardization methods rely on the standard assumptions of causal inference but differ in modeling. In IPW, we model the treatment D, whereas, for standardization, we model the outcome Y. Moreover, we demand that the model for treatment D condition on a set of confounders Z is correctly specified for IPW while the outcome model, Y conditional on treatment D and confounders Z is correctly specified for the standardization approach in order to yield consistent causal estimates.

A doubly robust estimator combines both the treatment model and the outcome model and only requires correct specification of either one of the models. Under identifiability assumptions, a doubly robust estimator yields consistent estimates for causal effect when one of the two models is correctly specified (Hernán and Robins, 2020). The estimator remains robust if one (but not both) of the models is misspecified (Robins et al., 2007). In general, doubly-robust estimators are commonly preferred over singly-robust approaches in estimating causal effects (Van der Laan and Rose, 2018).

The doubly robust method which is also known as the augmented IPW (AIPW), involves specifying models for both the outcome and the exposure based on covariates. We model the outcome and the set of confounders Z across the different exposure groups. Using the estimates from these models, we predict the outcomes $\hat{\mu}_1(Z_i)$ and $\hat{\mu}_0(Z_i)$ for each individual i under two different exposure conditions, either exposed ($D = 1$) or not exposed ($D = 0$), based on their covariate values Z. Additionally, we model the exposure as a function of the covariates to calculate the predicted probabilities $\hat{\pi}_0(Z_i)$ and $\hat{\pi}_1(Z_i)$ of being (un)exposed respectively for each individual.

We then estimate the counterfactual mean outcomes separately in the exposure groups using Equations 5 and 6 as described in (Hernan and Robins, 2020).

$$\hat{E}\left[Y^{(1)}\right]_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1(Z_i) + \frac{D_i}{\hat{\pi}_1(Z_i)}\left(Y_i - \hat{\mu}_1(Z_i)\right)\right) \tag{5}$$

$$\hat{E}\left[Y^{(0)}\right]_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_0(Z_i) + \frac{1-D_i}{\hat{\pi}_0(Z_i)}\left(Y_i - \hat{\mu}_0(Z_i)\right)\right) \tag{6}$$

The average causal effect can be computed by contrasting the estimated counterfactual means by taking the ratio in Equation 1. There are several doubly robust estimators suggested in the literature for estimating treatment effects. For example, machine learning methods can be incorporated into the doubly robust estimation to obtain doubly robust machine learning estimators. In the doubly robust estimator described above, we have seen that we need to obtain estimates $\hat{\mu}_d(z)$ and $\hat{\pi}_d(z)$; $d \in \{0,1\}$ (referred to as nuisance parameters) via the sort of traditional parametric models. However, these models are prone to misspecification, especially in high-dimensional settings. Thus, predictive machine learning algorithms such as tree-based algorithms can be adopted to estimate these conditional expectations. The performance of such estimators can be improved by cross-fitting, which removes the potential overfitting issues. In this study, we explore the doubly robust methods proposed by Yadlowsky et al. (2021) which are related but distinct from the AIPW. Their approach is optimized for estimating ratio based CATE scores for count outcomes and integrates the use of flexible machine learning estimators of the nuisance parameters while providing valid inference. They proposed contrast regression approach and provided another approach known as two regression approach.

Contrast Regression Approach

In this approach, the doubly robust estimator is constructed based on a semi-parametric model of treatment-covariate interactions building from the work of Van der Laan and Rose (2011), and Robins and Rotnitzky (2001) on semi-parametric models. Specifically, the treatment contrast in Equation 2 is estimated in the semi-parametric model $B(z) = \exp(\delta^T \mathbf{z})$, in which the conditional means $\mu_d(z) = E[Y^{(d)} \mid Z = z]$, and the propensity score, $\pi_d(z) = Pr(D = d \mid Z = z)$, are in non-parametric models. The CATE score is thus estimated from;

$$\hat{B}(z) = \exp(\hat{\delta}^T \tilde{z}),$$

where $\tilde{z}$ is z with an intercept i.e, $\tilde{z} = (1, z^T)^T$ and $\hat{\delta}$ is the estimate obtained by solving the doubly robust estimating equation described in Equation 7 in the Appendix.

Two Regression Approach

In this approach, separate regressions are fitted for each treatment group, adjusting for confounders. This allows interpretation of the regression models in each arm as if it were a randomized clinical trial. It was shown to yield a consistent estimate when there is no treatment heterogeneity (Yadlowsky et al., 2021). The CATE estimate score is obtained from;

$$\hat{B}_1(z) = \exp\left((\hat{\beta}_1 - \hat{\beta}_0)^T \tilde{z}\right),$$

where estimates $\hat{\beta}_1$ and $\hat{\beta}_0$ are obtained by fitting a poisson regression in the exposure groups, and solving estimating Equation 9 described in the Appendix.

A detailed procedure for the steps entailed for each of the approaches is outlined in the Appendix under Methods section and more details are discussed in Yadlowsky et al. (2021).

2.4.3 Bootstrap

Bootstrapping is a statistical resampling procedure that involves drawing random samples repeatedly from the original dataset to create many simulated samples. Based on the resampled data, one can approximate the distribution of the population from which the samples were obtained. The model-based standard errors obtained based on the casual inference estimator do not take into account the necessary steps performed while attaining balance for confounders between treatment groups. In order to make valid inferences, a procedure that takes into account these steps in variance estimation is desired. One way is to employ non-parametric bootstrap resampling procedures to compute the standard error, confidence intervals and p-values. In non-parametric bootstrap, we do not make any assumptions about the form of the underlying population distribution.

2.5 Software and Testing

All analyses were performed using R software version 4.2.3 (R Core Team, 2023). Multiplicity correction was done by adjusting p-values using the Benjamini Hochberg procedure (Benjamini and Hochberg, 1995), which controls for False Discovery Rate (FDR). All statistical tests were performed at 5% FDR level.

3 Results

3.1 Data Exploration

3.1.1 Descriptive statistics

Table 2 shows descriptive statistics of the covariates for the lupus cases and healthy individuals. From Table 2, the diseased and healthy individuals are different. For example, the diseased tend to be older (25 vs 18 years) than the healthy individuals.

Table 2: Distribution of the variables across SLE status. The Mean (standard deviation) for continuous covariate and number (proportion) for categorical covariates.

	Covariate	SLE (n = 158)	Healthy (n = 98)	Subtotal
Age:		24.8 (13.6)	18.1 (14.4)	
Sex:	Male	15 (88.2%)	2 (11.7%)	17
	Female	143 (59.8%)	96 (40.2%)	239
Pop_Cov:	Asian	83 (77.6%)	24 (22.4%)	107
	European	75 (50.3%)	74 49.7%)	149

3.1.2 Unsupervised data exploration

The underlying relationships among the sample groups were explored using Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) in order to visaulize the dataset in a low dimension by disease status. The plots in Figure 3 suggest some degree of clustering based among individuals on SLE status. The UMAP plot shows a clearer separation between the two groups compared to PCA, although there is still some overlap.

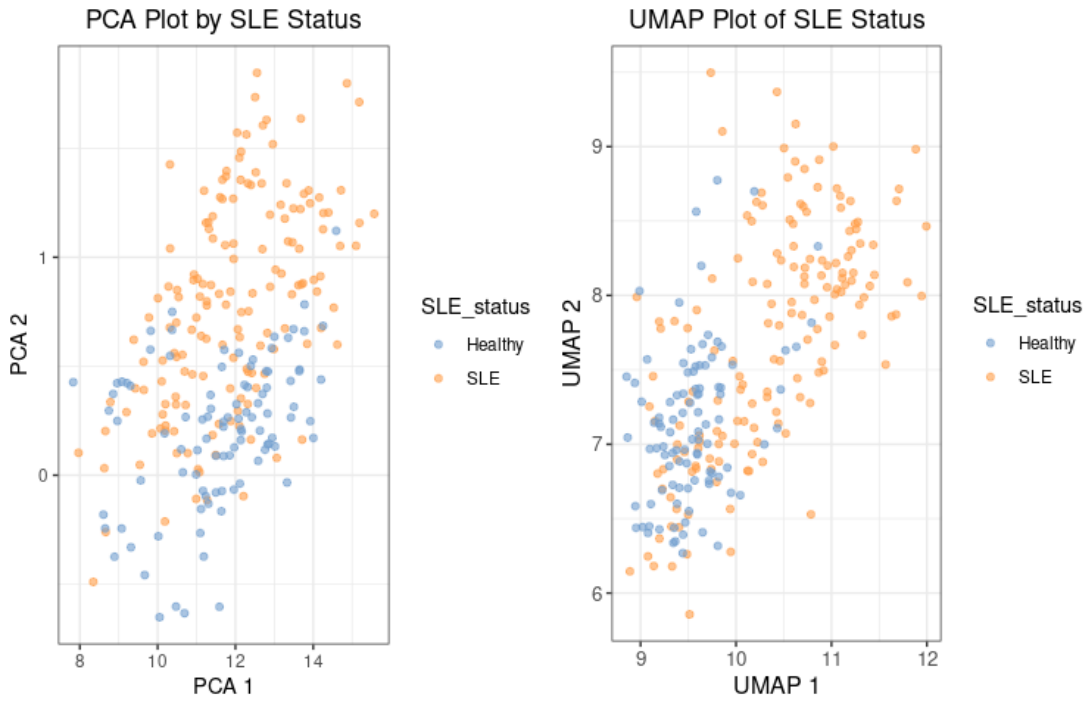


Figure 3: The PCA plot (on the left) and UMAP plot (on the right) visualize individuals by disease status in a 2-dimensional space.

3.2 Gene Filtering

Out of 12,869 genes, 8578 genes (about 67%) remained after gene filtering. We therefore considered 256 pseudo-bulk samples and 8578 genes for the gene expression analysis and exploration of the disease effect.

3.3 Conventional DGE Analysis

The traditional DGE analysis, where we simply associate the disease effect to gene expression was performed twice using gene by gene NB-GLMs with a log-link to evaluate the impact of confounding of the disease effect on the gene expression levels. We fitted the GLM models defined in Equations 10 and 11 in the Appendix using `glmmTMB` R package.

Naive Analysis:

Gene-by-gene GLM Models were fitted each with a linear predictor consisting of only the disease effect, i.e., Gene expression $\sim$ disease with Log-transformed normalized library size included as an offset. We discovered **5193 significant genes** at the 5% FDR level.

Table 3: Top 10 differentially expressed genes based on the Naive Analysis

Gene	EffectSize(log ratio)	StdError	Adjusted Pvalue
EPSTI1	1.4708	0.0754	9.46e-81
IFI27	3.3220	0.1904	1.47e-64
IFI6	1.5050	0.0951	6.28e-53
IFI44L	1.7718	0.1136	1.48e-51
SIGLEC1	1.7422	0.1150	1.30e-48
LGALS3BP	1.6235	0.1083	1.22e-47
IFI44	1.1802	0.0790	2.17e-47
ANXA2	0.3699	0.0248	2.90e-47
PSMA7	0.2804	0.0188	3.46e-47
ISG15	1.3938	0.0938	5.60e-47

Table 3 shows the effect size estimates for the top 10 differentially expressed genes presented as log fold changes. A fold change is interpreted as the ratio of the expected count in the diseased group compared to the expected count in the healthy group but with the same library size across the samples.

Conditional Analysis:

In this case, we fitted gene-by-gene GLM models with a linear predictor consisting of the disease effect and covariates `Age`, `batch_cov`, `Sex` and `PopCov`, i.e,

Gene expression $\sim$ disease + batch_cov + Age + Sex + PopCov. Like in the Naive case, log-transformed normalized library size was included in the model as an offset.

3269 genes were found to be significantly differentially expressed at 5% FDR level. Table 4 shows the effect sizes estimates presented as log fold changes for the top 10 differentially expressed genes where a fold change is the ratio of the expected gene expression in the diseased group as compared to the expected expression in the healthy group, adjusted for library size.

Table 4: Top 10 differentially expressed genes based on the Conditional Analysis

Gene	EffectSize(log ratio)	StdError	Adjusted Pvalue
MIR24-2	-0.7320	0.0519	3.03e-41
EPSTI1	1.2656	0.0959	4.22e-36
IFI27	2.9571	0.2259	1.06e-35
LGALS3BP	1.6349	0.1359	5.18e-30
MYL12A	0.2500	0.0209	1.10e-29
IFIT3	1.6895	0.1447	2.34e-28
IFIT1	1.7463	0.1520	1.87e-27
CHMP5	0.5492	0.0479	1.88e-27
OAS1	0.9407	0.0834	1.47e-26
IFI44L	1.7254	0.1533	1.87e-26

Comparison between the naive and conditional analyses

The results of the naive and conditional analyzes showed several key differences primarily due to the complexity of the models used. The conditional analysis identified fewer significant genes as compared to the naive analysis (3269 Vs 5193) genes being flagged at the same FDR of 5%. **2489 genes** were consistently identified as significant in both cases as shown in the venn diagram in Figure 4. The volcano plots in Figure 5 show that, for most of the genes in the naive case cluster around zero fold change whereas in the conditional case, there is somewhat a greater spread in the fold changes. The scatter plot in Figure 6a shows that for most of the genes, the effect sizes based on the naive analysis differ from those from the conditional analysis. This is indicated by the deviation of the points from the red dashed line. From Figure 6b, the conditional analysis tends to produce larger p-values (smaller values of $-\log_{10}$ p-values) as compared to the naive case since most of the points fall below the red dashed line.

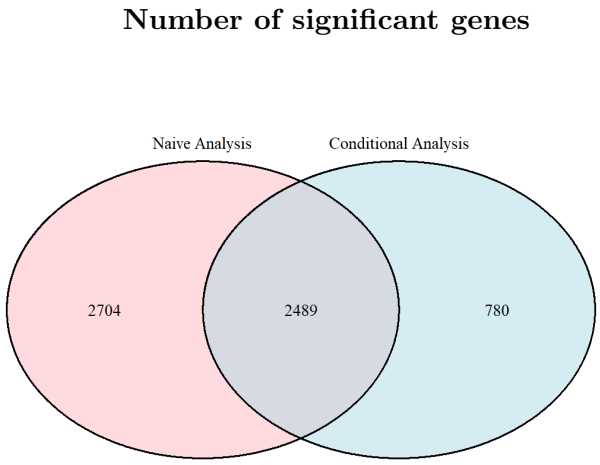


Figure 4: Venn diagram comparing the number of significant genes based on the naive analysis versus conditional analysis.

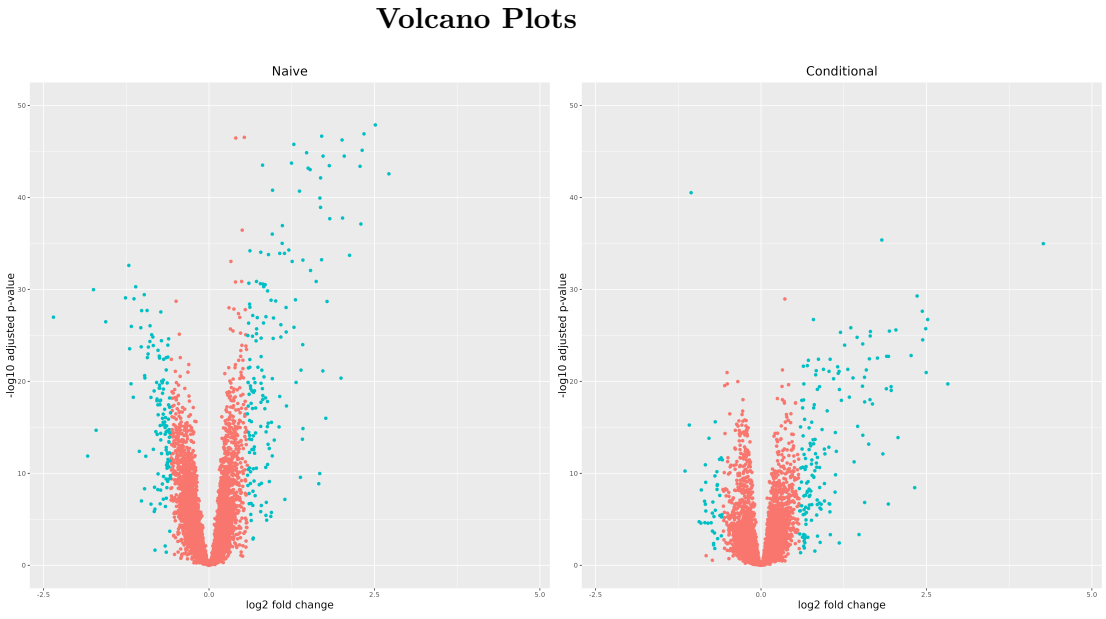
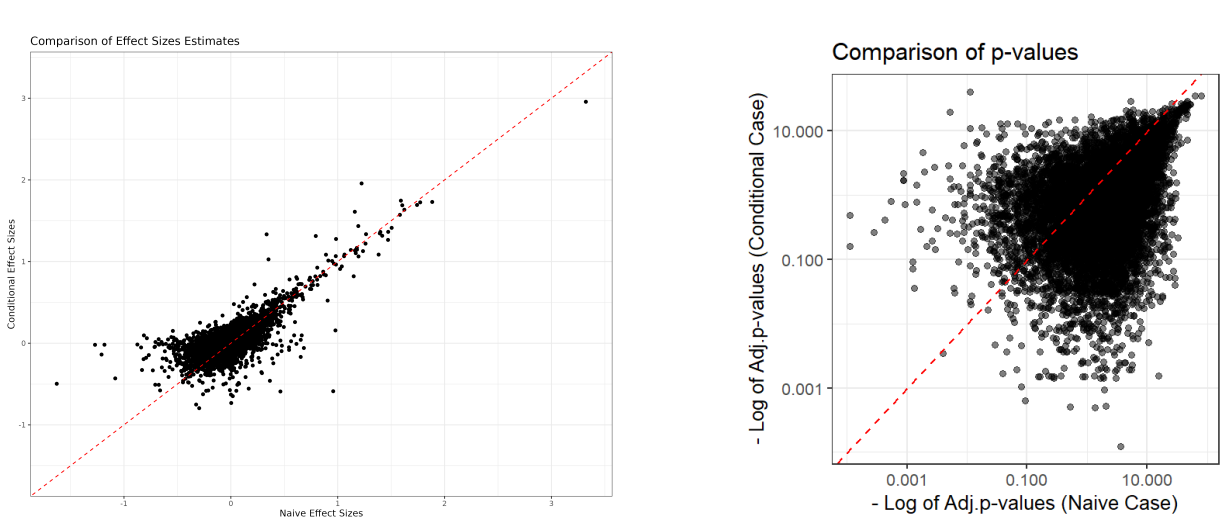


Figure 5: Volcano plots for the naive analysis versus conditional analysis.

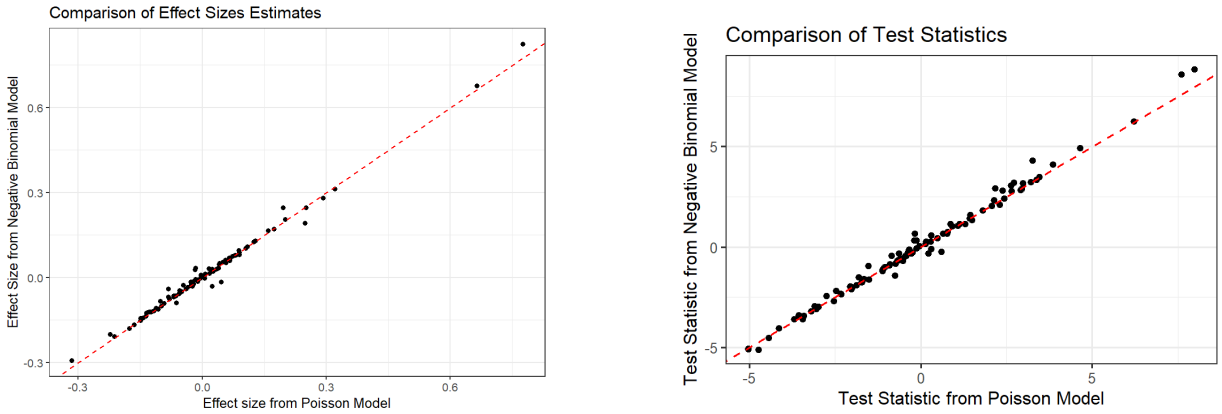


(a) Scatter plot for effect size estimates

(b) Scatter plot for p-values

Figure 6: Comparison of effect size estimates and p-values in the naive versus conditional analyzes

We explored the differences in the effect size estimates between using a Poisson-based gene by gene models and Negative Binomial models on a randomly sampled set of 100 genes. Figure 7a indicates a strong agreement between the effect sizes estimates for most of the genes, although slightly different for some genes. In addition, Figure 7b shows that the test statistic values from both models are also very close. Since bootstrap standard errors were used in constructing the test statistic values plotted in Figure 7b, these account for the mis-modeling of the variance in Poisson versus the negative binomial. This means that even if the fold changes (effect sizes) might be a little bit different for some genes, it could be that under repeated resampling, the distribution of the test statistic is actually very similar based on the bootstrap procedure which in turn result into similar inference irrespective of the model used. In fact from Figure 15 in the Appendix, the estimated dispersion parameters, k were relatively large (greater than 100) for most of the genes in the sampled set which implies relatively small values overdispersion ($\phi = \frac{1}{k}$) based on the assumed mean-variance relationship, variance $= \lambda + \frac{\lambda^2}{k}$ by `nbinom2` option in the `glmmTMB` package.



(a) Effect size estimates

(b) Test statistics

Figure 7: Scatter plots comparing effect size estimates and test statistic values based on the Poisson model versus the negative binomial model. The test statistics are computed based on bootstrap standard errors (bootSE), i.e., test statistic $= \frac{\text{effect size}}{\text{bootSE}}$.

Thus, for estimation of causal effects, we opted for the Poisson model to model the outcome due to its computational efficiency. This was due to the slower fitting times observed with negative binomial models, which, although they may handle overdispersion better, required significantly more computational resources and time, making them less practical for our analysis needs. Fitting based on all the genes based on Poisson regression including all the covariates yielded **4317 significant genes** at 5% FDR level.

3.4 Causal Effect Estimation

In order to estimate the causal disease effect, standard causal methods standardization, Inverse probability weighting and doubly robust estimation approaches were applied to estimate the ratio based marginal average causal effect and Yadlowsky approaches to estimate the ratio based CATE.

3.4.1 Standardization Approach

In order to estimate the causal effect based on standardization, we fit Poisson GLMs to regress gene expression (outcome) over confounders using the `glm()` function with a log link, separately for diseased and healthy groups (see Equation 12 in the Appendix). The fitted models were then used to predict the expression levels in each group for every individual, corresponding to the two

estimated average potential outcomes. These were then averaged over all individuals separately in each exposure group. The causal disease effect was thus computed by taking the ratio of the estimated average potential outcomes. Non-parametric bootstrap resampling procedure (with replacement) was used to estimate the standard error and the p-values using 200 bootstrap replicates. Table 5 shows the top 10 significant genes based on the bootstrap adjusted p-values. The effect sizes shown in Table 5 are the log-transformed ratios which are interpreted as the log-ratio of the expected counts if all individuals would be diseased versus if all individuals would be healthy.

Table 5: Top 10 differentially expressed genes based on the standardization method

Gene	EffectSize (log ratio)	StdError	Adjusted PValue
IFI27	2.9855	0.1967	8.57e-3
CXCL10	1.9707	0.1325	8.58e-3
MTRNR2L8	2.9616	0.2386	8.58e-3
SIGLEC1	1.7950	0.1554	8.59e-3
IFIT3	1.7374	0.1633	8.59e-3
IFIT1	1.7207	0.1542	8.59e-3
ANKRD22	1.6617	0.1542	9.00e-3
RSAD2	1.6302	0.1589	9.00e-3
IFI44L	1.6071	0.1423	9.00e-3
LGALS3BP	1.5727	0.1817	9.01e-3

Based on the standardization approach, **3701 genes** were significant at 5% FDR level.

3.4.2 Inverse Probability Weighting

Estimation of causal disease effect based on IPW involved fitting a logistic regression treatment model (see Equation 13 in Appendix) with a binary outcome given covariates Z. This was then used to obtain the probabilities of being diseased or healthy for each individual. The stabilized weights for each individual were computed as shown in Equation 4. The weighted regression outcome model (Equation 14 in Appendix) given treatment was then fitted to estimate the causal effect. Like in the standardization approach, standard errors and p-values were computed based on non-parametric bootstrap with 200 bootstrap replicates. Table 6 shows the top 10 significant genes based on the bootstrap adjusted p-values. The effect sizes shown in Table 6 are the log-transformed ratios which are interpreted as the log-ratio of the expected counts if all individuals would be diseased versus if all individuals would be healthy.

Table 6: Top 10 differentially expressed genes based on the IPW method

Gene	Effect Size (log ratio)	Std Error	Adjusted PValue
IFI27	3.3296	0.1995	1.17e-2
USP18	1.9194	0.1712	1.18e-2
IFI44L	1.8303	0.1152	1.18e-2
SIGLEC1	1.7839	0.1310	1.18e-2
LINC00910	-1.7510	0.2236	1.18e-2
LGALS3BP	1.6962	0.1082	1.19e-2
IFIT3	1.6879	0.1085	1.19e-2
IFIT1	1.6814	0.1151	1.19e-2
RSAD2	1.6482	0.1085	1.19e-2
IFITM3	1.5739	0.1029	1.21e-2

Based on the Inverse Probability Weighting approach, **3641 genes** were significant at 5% FDR level.

Comparison between Standardization and IPW Effect Estimates

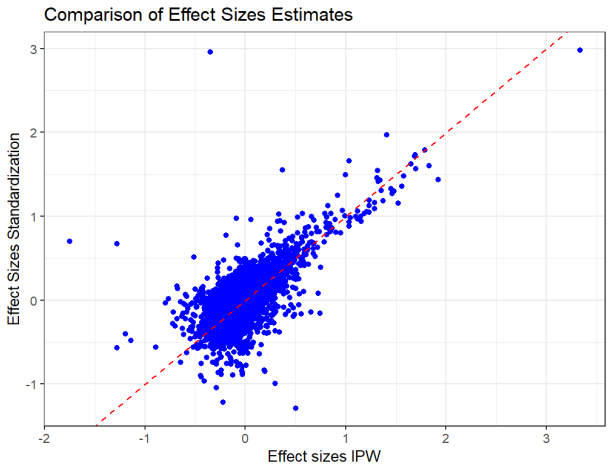


Figure 8: A scatter plot comparing effect size estimates based on Standardization and IPW. Each data point corresponds to a gene and the dashed red line represents the line of equality if the effect sizes estimated by both methods were the exactly the same. Points along this line indicate that both IPW and Standardization yielded the same effect size for those specific genes.

Comparing number of significant genes based on IPW versus Standardization

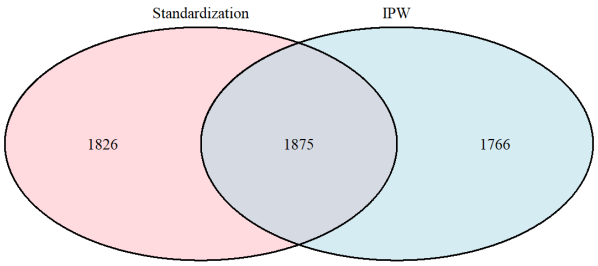


Figure 9: A Venn diagram comparing the number of significant genes based on IPW versus Standardization.

From Figure 8, most of the genes have their effect sizes clustered around the red dashed line and the points show a strong linear relationship between the two sets of estimates. However, For some genes, there is a deviation between the effect size estimates based on both methods, especially for larger effect sizes. The venn diagram in Figure 9 illustrates the number of significant genes identified by each method independently and those identified by both IPW and standardization. **1875 genes** were found significant at 5% FDR level by both methods. The substantial overlap suggests some agreement between the methods on a core set of genes. Additionally, the two methods in comparison identified unique sets of genes, nearly half ($\sim$ 50 %) of the respective total number of differentially expressed genes, highlighting differences between the methods. We further explored the degree of overlap among the top 100 differentially expressed genes in Figure 16 in the Appendix. Out of the 100 genes, 62 were discovered as differentially expressed by both methods and only 38 were uniquely identified by each method. This indicates that there is a substantial overlap (of more than 50%) in the top list.

Doubly Robust Approach

We further estimated the average treatment effect (ratio between the diseased versus non diseased) with a doubly robust estimator using `atefitcount()` implemented in precmed package (Tian et al., 2024). The outcome model and the propensity score model were specified as shown in Equations 15 and 16 respectively. The function computes the ratio of the estimates in Equations 5 and 6 which yield the average treatment effect between the diseased versus non diseased.

It also uses bootstrap resampling procedure for inference and 200 bootstrap samples were considered by setting `n.boot = 200`. The effect size estimates were then compared with IPW and Standardization approaches. Table 7 shows the top 10 differentially expressed genes based on the doubly robust approach. The effect sizes shown in Table 7 are the log-transformed ratios which are interpreted as, the log-ratio of the expected counts if all individuals would be diseased versus if all individuals would be healthy.

Table 7: Top 10 differentially expressed genes based on the AIPW - doubly robust approach

Gene	Effect Size (log-ratio)	Std Error	Adjusted Pvalues
IFI27	2.9676	0.3084	< 0.0001
SIGLEC1	1.7063	0.1788	< 0.0001
CMPK2	1.1657	0.1298	< 0.0001
OAS3	1.0857	0.1141	< 0.0001
OAS2	1.0455	0.0961	< 0.0001
OAS1	1.0195	0.1040	< 0.0001
MT2A	0.9888	0.1037	< 0.0001
SAMD9L	0.9754	0.1048	< 0.0001
XAF1	0.9371	0.1060	< 0.0001
RTP4	0.8974	0.1006	< 0.0001

2609 genes were discovered to be significant at 5% FDR level.

Comparison between doubly robust method with IPW and Standardization.
Generally, the two scatter plots in Figure 10 indicate a strong linear relationship between the effect sizes, with most points aligning closely along the red dashed line. However, there are also some deviations observed (especially for more extreme effect sizes). The variability is more pronounced in the standardized case versus doubly robust (right plot) as compared to IPW versus doubly robust (left plot). The UpSet plot in Figure 11 provides a comparative analysis of the significant findings identified by three estimation approaches: Doubly Robust (DR), Inverse Probability Weighting (IPW), and Standardization (STD). The largest intersection includes **1371 genes** identified by all three methods, highlighting consensus across different approaches. All three approaches discovered different number of significant genes, highlighting differences between the estimation approaches. The number of significant genes discovered by both IPW and the DR approach were more than those discovered by both STD and the DR approach (728 vs 352 genes). Additionally, the DR approach has the fewest (158 genes) unique findings indicating that most of the genes discovered as significant based on the DR approach were also significant in

either IPW, STD or both. Further, the Upset plot in Figure 17 in the Appendix, compares the significant findings based on DR, IPW, and STD on the top 100 differentially expressed genes. All the three methods identified 61 genes in common as significant. This shows strong agreement between the three approaches especially in the top list.

Comparison between doubly robust effect sizes with IPW and Standardization.

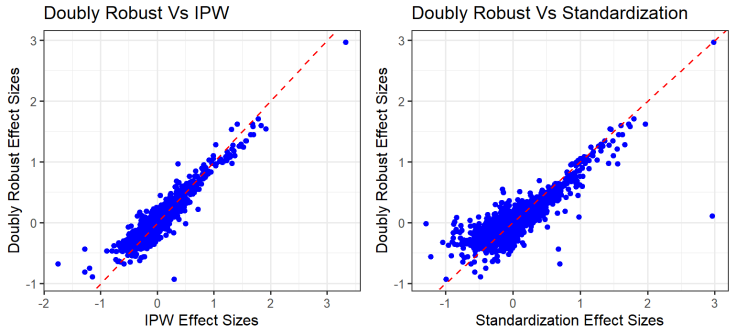


Figure 10: A scatter plot comparing effect size estimates based on doubly robust versus IPW (on the left) and doubly robust versus Standardization (on the right). Each data point is a gene and the dashed red line represents the line of equality of the effect sizes estimated by both methods in comparison. Points along this line indicate same effect size estimates based on the methods being compared for those specific genes.

Number of significant genes discovered.

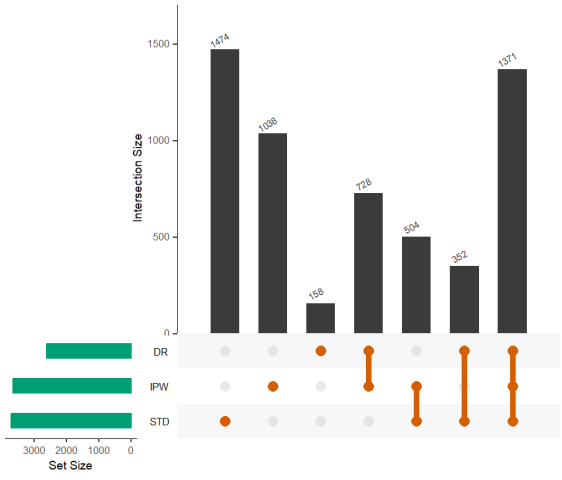


Figure 11: Upset plot comparing the number of significant genes based IPW, Standardization (STD) and Doubly robust (DR) estimation.

Yadlowsky Approaches

Yadlowsky et al. (2021) approaches were used to estimate the individualized ratio-based CATE score defined as; $\text{CATE}(z) = \log\left[\frac{E[Y^{(1)}|Z=z]}{E[Y^{(0)}|Z=z]}\right]$, which contrasts potential outcomes in the diseased group versus in the healthy group. $\text{CATE}(z)$ score indicates the individualized log ratio if an individual would be diseased ($D = 1$) versus if they would be healthy ($D = 0$), conditional on the baseline covariates. In order to implement Yadlowsky methodology, we estimated individualized CATE scores using the `catefitcount()` function implemented in the precmed R package (Tian et al., 2024). The score methods, **Contrast regression** and **Two regressions** were specified in the *score.method* argument as "contrastReg" and "twoReg" respectively, utilizing boosting-based prediction as the initial prediction.

Due to time constraints, we evaluated the two approaches on a selected set of only six genes because it took approximately six minutes to obtain the scores for all individuals per gene. Specifically, we selected six genes (among the significant genes based on the previous approaches) and estimated the ratio based individualized CATE scores on a log scale which is the log-ratio of the expected counts if a person were diseased over the expected counts if they were not diseased, conditional on the covariates.

The distribution of the scores for each of the selected genes estimated based on Contrast regression and Two regressions methods are visualized in Figures 12 and 13 respectively. In Figure 12, the estimated log transformed CATE scores for all/most of the individuals are greater than 0, for genes **IFI27**, **OAS1**, and **MT2A**. This means that that those genes are highly expressed if the persons would be diseased versus if they would be healthy. Whereas genes **KIF22**, **MT-ND4L**, and **EIF4B** in the same Figure, are highly expressed if the persons would be healthy versus if they would be diseased, since the distribution of their log CATE scores indicate that for most persons, the log CATE scores are less than 0 (negative).

For the same genes, similar trends and conclusions can be deduced from the distribution of the log CATE scores in Figure 13 estimated based on Two regression approach. This indicates agreement in the estimated log-CATE scores based on the two approaches.

Distribution of the log CATE scores estimated based on Contrast regression.

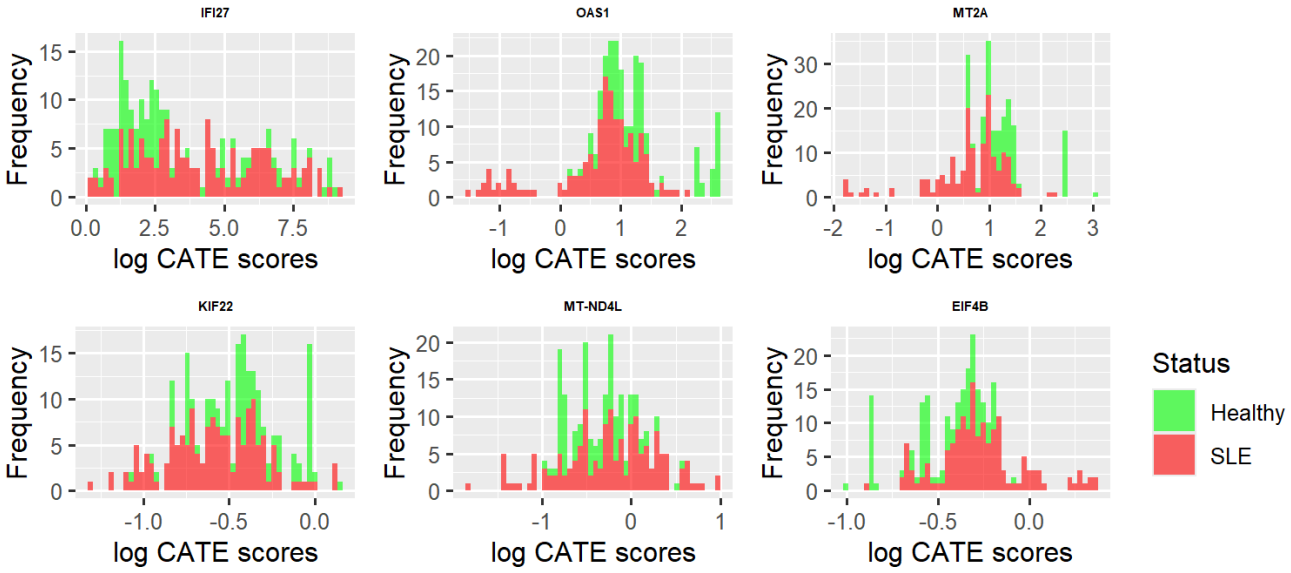


Figure 12: Stacked histograms showing the distribution of the log CATE scores for each of the selected genes computed based on Contrast regression score method. The green and the red bars indicate the healthy and diseased individuals respectively.

Distribution of the log CATE scores estimated based on Two regression.

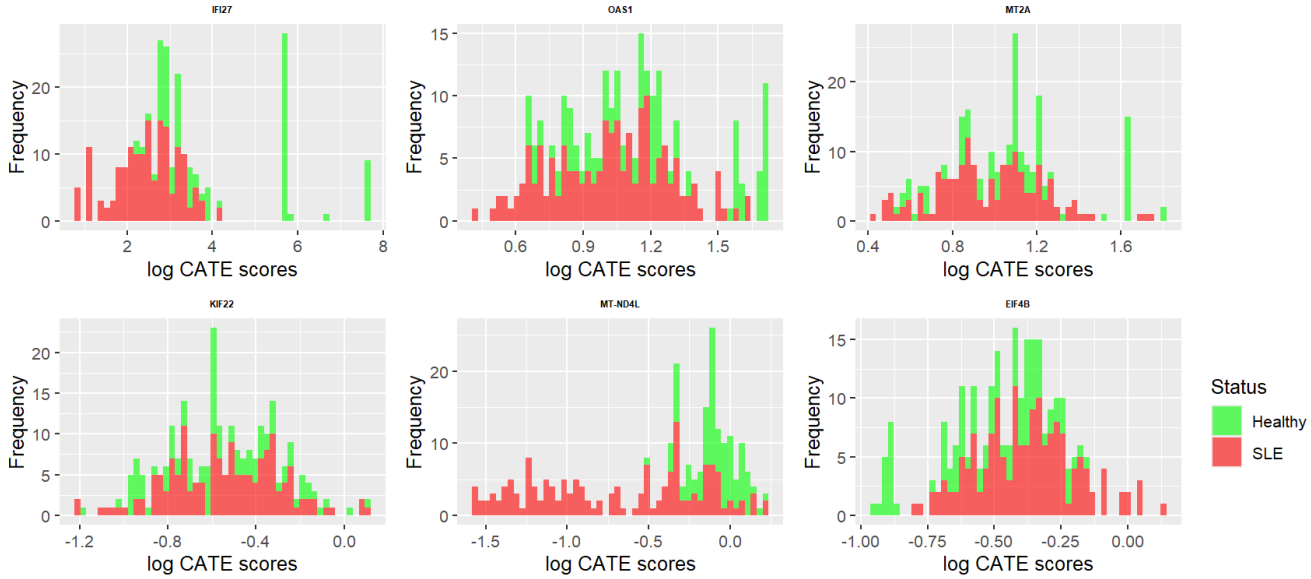


Figure 13: Stacked histograms showing the distribution of the log CATE scores for each of the selected genes computed based on Two regression score method.The green and the red bars indicate the healthy and diseased individuals respectively.

Figure 14 depicts the density plots comparing the log CATE scores estimated based on Contrast regression and Two regressions methods. From the plot, the log CATE scores span the same ranges for the same gene for the two approaches, and they both result into the same conclusions for atleast the evaluated genes.

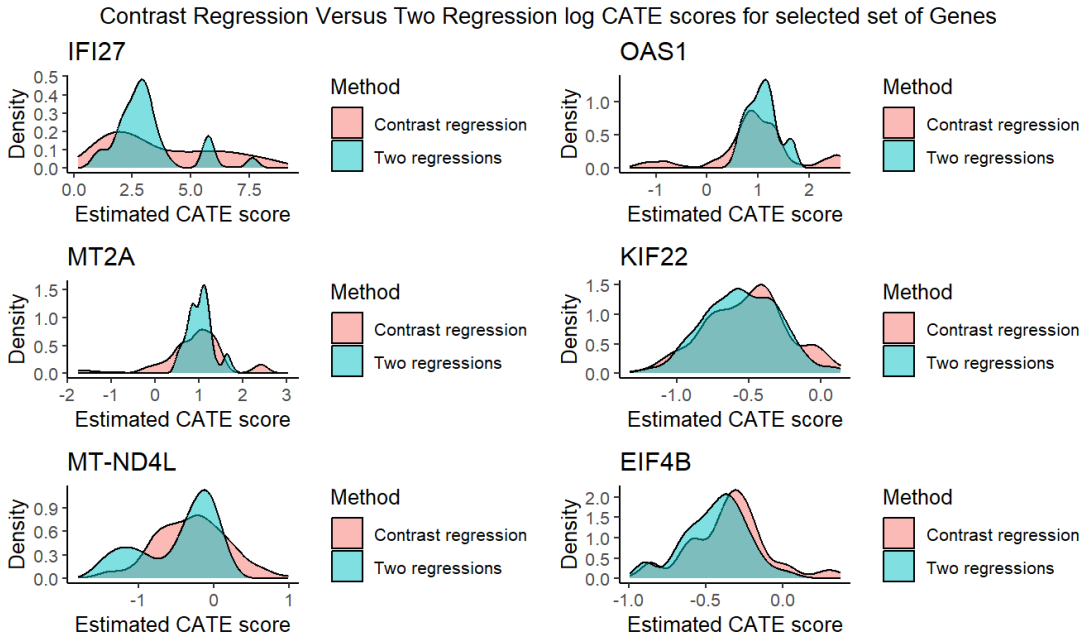


Figure 14: Density plots comparing the log CATE scores estimated based on Contrast regression versus Two regressions methods.

4 Discussion

While performing differential gene expression (DGE) analysis, it is often of interest to investigate effects directly caused by the exposure (treatment) on the gene expression and lists of affected genes. This can be achieved through proper adjustment for the potential confounders in the analysis. In this study, we performed causal DGE on a single cell - gene expression dataset comparing lupus patients versus healthy controls. Causal inference techniques; standard IPW, Standardization and doubly robust estimators were applied to estimate the ratio based average treatment effect (Diseased versus Healthy). The effect size estimates and the number of differentially expressed (DE) genes were thus compared. The individualized CATE scores were estimated using contrast regression and two regression methods proposed by Yadlowsky et al. (2021).

First, traditional differential gene expression analysis was performed twice; using only the disease status as the linear predictor (naive analysis) and adding disease status plus other covariates in the model (conditional analysis). This was to evaluate the impact of confounding of the disease effect the gene expression. The results showed notable differences in the number of significant genes, effect size estimates and p-values. For instance, the conditional analysis identified fewer significant genes as compared to the naive analysis (3269 Vs. 5193) being flagged at the same FDR of 5%. This can be attributed to the fact that inclusion of additional covariates might control for confounding factors, leading to low false discoveries in the conditional analysis and thus fewer significant discoveries as compared to the naive case. In fact, one would typically expect that adding covariates results in higher power and thus could result into more DE genes. This is not the case here. The possible reason is that the covariates are confounders and thus associated with the disease variable which results into multicollinearity issue, hence fewer DE genes in the conditional case.

Moreover, some genes like **EPSTI1** and **IF127** appeared in the top 10 DE genes for both analyses but with different effect size estimates and p-values, showcasing how additional covariates in the model can influence the estimated impact of the disease. It is worth noting that a straightforward analysis focusing solely on the disease effect may overestimate the true impact of the disease by ignoring potential confounders leading to biased results (Barrowman et al., 2019). Conversely, including additional covariates can provide more reliable estimates of the effect.

The Poisson-based and negative binomial effect size estimates evaluated on a randomly selected set of 100 genes, indicated a strong agreement between the effect sizes estimates for most of the genes although slightly different for some genes. The results from the comparison between the test statistic values constructed based on the bootstrap standard errors could result in similar inferences irrespective of the model used. Recognizing the fact that negative binomial models would handle over-dispersion better, it required significantly more computational resources and time, making them less practical for our analysis needs. We, therefore, opted for the Poisson model to model the outcome in causal effect estimation due to the slower fitting times observed with negative binomial models.

On comparison between IPW and Standardization approaches in Figure 8, the effect size estimates were similar for most of the genes. Further, there is a strong linear relationship between the two sets of estimates, indicate that even when the estimates are not exactly the same, they tend to increase or decrease together. For some genes, a clear deviation of points from the red dashed diagonal line was observed, indicating differences in the effect sizes as estimated by the two approaches. The differences are also manifested in the number of uniquely identified genes.

While IPW and standardization target the same estimand (average treatment effect) and aim to achieve the same goal, the approaches have got different ways of estimating the effect and require slightly different assumptions. The IPW estimator depends on accurately specifying the exposure model, which is generally more feasible than accurately specifying the outcome model, which is desired in the standardization approach (Reifeis et al., 2020). However, IPW may be vulnerable to the influence of extreme weights. Thus, the differences observed in the effect size estimates are indeed expected, even worse if there is strong confounding. It is also worth noting that the effect of confounding may not be the same on every gene. Regarding the number of significant genes, IPW discovered fewer differentially expressed genes at 5% FDR level as compared to standardization (3641 Vs 3701). In general, there are minor differences in bias and efficiency between the IPW and standardization estimators, and one should choose between the two approaches based on their knowledge of how well the exposure or outcome models are specified.

The doubly robust approach also yielded comparable estimates as those based on IPW and standardization, although with a lower number of significant genes - **2609 genes** as compared to the singly robust methods. The effect size estimates in the standardized case versus doubly robust were more variable as compared to IPW versus doubly robust. Further, the number of significant genes discovered by both IPW and the doubly robust approach were more than those discovered by both standardization and the doubly robust approach (728 vs 352 genes). These comparisons indicate that the IPW approach seems to be closer to doubly robust, perhaps suggesting that the propensity model may be better fitting the data than the outcome model.

In practice, we rarely know the true relations among exposure, outcome, and confounders. Doubly robust estimators have shown to provide advantages over IPW and standardization approaches (Moodie et al., 2018), since they provide two chances to get it right i.e, yield consistent estimates for causal effect when atleast one of the two models is correctly specified and thus often preferred. It is also worth noting that the three causal effect methods discovered lower number significant genes as compared to the traditional Poisson based method.

Yadlowsky et al. (2021) approaches were implemented to estimate the ratio-based CATE scores. Their approaches focus on estimating "individualized" ratio-based CATE scores utilizing doubly robust estimation and semi-parametric modeling. It also integrates flexible machine learning methods such as boosting in the predictions. Due to computational time, the scores were estimated and evaluated on a selected set of six genes. For the same genes, the estimated log-CATE scores based on the Contrast regression and the Two regression approaches in Figures 12 and 13 respectively, yield similar conclusions.
Further, conclusions from the marginal log-ratio estimates for the selected genes indicated some similarity with the conclusions based on the individualized conditional log-ratios; For instance, the AIPW flagged genes **IFI27**, **OAS1**, and **MT2A** as up-regulated and genes **KIF22**, **MT-ND4L**, and **EIF4B** as down-regulated. Note that we cannot directly compare the marginal estimands and the conditional estimands. The log CATE scores based on the two proposed approaches span the same ranges for the same gene and both result into the same conclusions for atleast the evaluated genes. In order to select which of the score method captures the most treatment heterogeneity (performs well), cross validated CATE scores are considered to reduce optimism.

5 Ethical Thinking, Societal Relevance, and Stakeholder Awareness

5.1 Ethical Thinking

Ethical considerations include ensuring data privacy and confidentiality, maintaining integrity, and fostering transparency throughout all research phases. Compliance with these standards is crucial for ethically handling patient data and ensuring the trustworthiness of study results. The project commits to adhering to the thesis agreement, which includes protocols for managing sensitive data and mandates accurate and complete reporting of findings to prevent misinterpretation and misuse.

5.2 Societal Relevance

The study aims to enhance treatment strategies for lupus through advanced causal inference techniques applied to genomic data, and it seeks to advance the methodologies of causal inference itself. By improving treatment outcomes, the project directly contributes to better patient care and quality of life. Additionally, by exploring several methodologies in causal inference, the project contributes to broader scientific and medical advancements. These contributions are a key for healthcare decision-making, offering long-term benefits to the medical community and patients alike by fostering methodological innovation and practical healthcare improvements.

5.3 Stakeholder Awareness

Key stakeholders in this project include Janssen Pharmaceutica, medical researchers, and lupus patients. Janssen Pharmaceutica benefits from enhanced drug development processes and insights into disease mechanisms. Medical researchers gain from improved analytical methods and enhanced scientific knowledge, which in turn inform clinical decision-making and target new therapeutic avenues. Patients stand to gain the most directly, as advancements in treatment offer them better care options and an improved understanding of their condition. Awareness of these stakeholders' perspectives ensures that the project's objectives align with the needs and expectations of those it aims to serve, promoting a more targeted and beneficial outcome.

6 Conclusions and Future Research

The study results have demonstrated that confounding factors have a huge impact on the exposure effect estimates as well as the number of differentially expressed genes and thus, proper adjustment should of these confounding variables should not be overlooked especially in observational studies. Further, while estimating causal effects of the exposure variable on the outcome, the estimator really matters even though they might be targeting the same estimand. Choosing between which estimator to use, for example between IPW and standardization, one can base on their relative confidence in specification of the exposure or outcome models.

Doubly robust estimators provide advantages over IPW and standardization and potentially allow the easing of certain modeling assumptions in observational genomics. Hence, preferred over singly robust estimators. Individualized ratio-based CATE scores were estimated based on Yadlowsky's doubly robust methods, Contrast regression, and the Two regressions approaches. These scores show effect heterogeneity among individuals based on different covariate levels. The individualized effects allow clinicians to understand how much an individual would benefit from a particular intervention based on the covariates. It should be noted that causal inference from an observational study cannot replace causal inference from an RCT. Conditional exchangeability is an untestable assumption in an observational study, whereas in an RCT, it holds true in practice.

The study opens for a wide range of opportunities in exploring ratio based CATE as a measure of causal effect. Future work may focus on how the individualized scores can be averaged across all individuals in order to perform inference on a gene. The study can also be extended to a setting with a continuous or longitudinal exposure variables.

References

Altman, N., & Krzywinski, M. (2015). Points of significance: Association, correlation and causation. *Nature methods*, *12*(10).

Andrews, T. S., Kiselev, V. Y., McCarthy, D., & Hemberg, M. (2021). Tutorial: Guidelines for the computational analysis of single-cell rna sequencing data. *Nature protocols*, *16*(1), 1–9.

Barrowman, M. A., Peek, N., Lambie, M., Martin, G. P., & Sperrin, M. (2019). How unmeasured confounding in a competing risks setting can affect treatment effect estimates in observational studies. *BMC Medical Research Methodology*, *19*, 1–11.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289–300.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). Glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, *9*(2), 378–400.

Carter, E. E., Barr, S. G., & Clarke, A. E. (2016). The global burden of sle: Prevalence, health disparities and socioeconomic impact. *Nature reviews rheumatology*, *12*(10), 605–620.

Crowell, H. L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., & Robinson, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature communications*, *11*(1), 6077.

Du, J.-H., Zeng, Z., Kennedy, E. H., Wasserman, L., & Roeder, K. (2024). Causal inference for genomic data with multiple heterogeneous outcomes. *arXiv preprint arXiv:2404.09119*.

Dukes, O., & Vansteelandt, S. (2018). A note on g-estimation of causal risk ratios. *American journal of epidemiology*, *187*(5), 1079–1084.

Goetghebeur, E., le Cessie, S., De Stavola, B., Moodie, E. E., Waernbaum, I., & the topic group Causal Inference (TG7) of the STRATOS initiative, ". (2020). Formulating causal questions and principled statistical answers. *Statistics in medicine*, *39*(30), 4922–4948.

Gohlmann, H., & Talloen, W. (2009). *Gene expression studies using affymetrix microarrays*. Chapman; Hall/CRC.

Hejazi, N. S., Boileau, P., van der Laan, M. J., & Hubbard, A. E. (2023). A generalization of moderated statistics to data adaptive semiparametric estimation in high-dimensional biology. *Statistical Methods in Medical Research*, *32*(3), 539–554.

Hernan, M., & Robins, J. (2020). Causal inference: What if. boca raton: Chapman & hill/crc.

Hernán, M. A., & Robins, J. M. (2020). Causal inference: What if.

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, *24*(8), 550–572.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, *47*(260), 663–685.

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Kotliar, D., Veres, A., Nagy, M. A., Tabrizi, S., Hodis, E., Melton, D. A., & Sabeti, P. C. (2019). Identifying gene expression programs of cell-type identity and cellular activity with single-cell rna-seq. *Elife*, *8*, e43803.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, *116*(10), 4156–4165.

Lee, S., & Lee, W. (2022). Application of standardization for causal inference in observational studies: A step-by-step tutorial for analysis using r software. *Journal of Preventive Medicine and Public Health*, *55*(2), 116.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Álvarez-Varela, A., Batlle, E., Sagar, n., Gruen, D., Lau, J. K., et al. (2020). Benchmarking single-cell rna-sequencing protocols for cell atlas projects. *Nature biotechnology*, *38*(6), 747–755.

Moodie, E. E., Saarela, O., & Stephens, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, *7*(1), e205.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, 1–51.

Perez, R. K., Gordon, M. G., Subramaniam, M., Kim, M. C., Hartoularos, G. C., Targ, S., Sun, Y., Ogorodnikov, A., Bueno, R., Lu, A., et al. (2022). Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, *376*(6589), eabf1970.

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reifeis, S. A., Hudgens, M. G., Civelek, M., Mohlke, K. L., & Love, M. I. (2020). Assessing exposure effects on gene expression. *Genetic epidemiology*, *44*(6), 601–610.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, *7*(9-12), 1393–1512.

Robins, J., & Rotnitzky, A. (2001). Analysis of a randomized trial with non-compliance and a binary outcome. *Proceedings of the 53rd International Statistical Institute meeting, Seoul, Korea*.

Robins, J., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statistical Science*, *22*(4), 544–559.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, *11*, 1–9.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, *66*(5), 688.

Smith, M. J., Mansournia, M. A., Maringe, C., Zivich, P. N., Cole, S. R., Leyrat, C., Belot, A., Rachet, B., & Luque-Fernandez, M. A. (2022). Introduction to computational causal inference using reproducible stata, r, and python code: A tutorial. *Statistics in medicine*, *41*(2), 407–432.

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J., Barraud, Q., et al. (2021). Confronting false discoveries in single-cell differential expression. *Nature communications*, *12*(1), 5692.

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, *16*(3), 133–145.

Tian, L., Jiang, X., & Simoneau, G. (2024). *Precmed: Precision medicine* [R package version 1.0.0.9000]. https://smartdata-analysis-and-statistics.github.io/precmed/

Van der Laan, M. J., & Rose, S. (2011). *Targeted learning* (Vol. 1). Springer.

Van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science*. Springer.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Xie, Y., Brand, J. E., & Jann, B. (2012). Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, *42*(1), 314–347.

Yadlowsky, S., Pellegrini, F., Lionetto, F., Braune, S., & Tian, L. (2021). Estimation and validation of ratio-based conditional average treatment effects using observational data. *Journal of the American Statistical Association*, *116*(533), 335–352.

Appendix

Methods:

We describe the steps for estimating the ratio-based CATE based on the approaches proposed by Yadlowsky et al. (2021). Let;

- D denote binary treatment variable taking values in the set $\{0, 1\}$.
- Z is a vector of baseline covariates.
- Y is a count outcome.

The aim is to estimate the ratio-based CATE score defined as; $\text{CATE}(z) = \log\left[\frac{E[Y^{(1)}|Z=z]}{E[Y^{(0)}|Z=z]}\right]$ where $Y^{(d)}$ is the potential outcome that would be observed if the patient received the treatment $d \in \{0, 1\}$. CATE(z) score indicates the individualized log ratio if an individual would be treated ($D = 1$) versus if they would be untreated ($D = 0$), conditional on the baseline covariates.

Contrast Regression Approach

`Steps`

1. Divide dataset randomly R into K non-overlapping parts, $R_1, R_2, \ldots, R_K$, of approximately equal size.
2. For each fold $k = 1, \ldots, K$ and each treatment group $d \in \{0, 1\}$:
   (a) Estimate the conditional mean outcome given the baseline covariates using a Poisson-based gradient boosting regression method based on the observations without the k^{th} fold, R_{-k}. Denote this prediction as $\hat{Y}^{(d)}_{-k}(z)$.
   (b) Estimate the Propensity Score (PS) on R_{-k}. Denote the estimated PS as $\hat{\pi}_{-k}(z)$.
3. Solve for δ in the doubly robust estimating equation; Let $\tilde{z}$ denote the z with an intercept,

$$S(\delta) = \sum_{k=1}^{K} \sum_{i \in R_{-k}} \tilde{z}_i \left[\frac{D_i \left\{ Y_i - \frac{1}{2}\left(\hat{Y}^{(0)}_{-k}(\tilde{z}_i) \exp(\delta^T \tilde{z}_i) + \hat{Y}^{(1)}_{-k}(\tilde{z}_i) \right) \right\} (1 - \hat{\pi}_{-k}(\tilde{z}_i))}{\exp(\delta^T \tilde{z}_i)\hat{\pi}_{-k}(\tilde{z}_i) + 1 - \hat{\pi}_k(\tilde{x})} \right.$$
$$\left. + \frac{(1 - D_i) \left\{ Y_i - \frac{1}{2}\left(\hat{Y}^{(0)}_{-k}(\tilde{z}_i) + \hat{Y}^{(1)}_{-k}(\tilde{z}_i) \exp(-\delta^T \tilde{z}_i) \right) \right\} \exp(\delta^T \tilde{z}_i)\hat{\pi}_{-k}(\tilde{z}_i)}{\exp(\delta^T \tilde{z}_i)\hat{\pi}_{-k}(\tilde{z}_i) + 1 - \hat{\pi}_{-k}(\tilde{z}_i)} \right] = 0, \quad (7)$$

   using the Newton-Raphson method (NRM) or an L2-norm score method as alternative in case of convergence issues with NRM. Denote the estimator as $\hat{\delta}$.

4. Repeat the above steps for B bootstrap samples to obtain $\hat{\delta}^b$ in the b sample, and compute the final estimator $\hat{\delta}$ as the mean of the bootstrap estimates $\hat{\delta}^b$.

5. Estimate the Conditional Average Treatment Effect (CATE) with contrast regression from:

$$\hat{CATE}_{\text{contrastreg}}(z) = \hat{\delta}^\top \tilde{z}.$$

Two Regressions Approach

`Steps`

1. Divide randomly the dataset R into K non-overlapping folds (subsets) of approximately equal size, labeled as $R_1, R_2, \ldots, R_K$.

2. For each fold k and each treatment group $d \in \{0, 1\}$:

   (a) Exclude the k-th fold to form R_{-k} (data excluding fold k). Then use a Poisson-based gradient boosting regression on R_{-k} to predict outcomes, adjusting for covariates. The predictions are denoted as $\hat{Y}_{-k}^{(d)}(z)$ which are the initial nonparametric prediction of the potential outcome.

   (b) Estimate the PS for each individual based on R_{-k}. Denote the PS as $\hat{\pi}_{-k}(z)$ and estimate the weights $\hat{W}^{(d)} = d\frac{D}{1-\hat{\pi}_{-k}(z)} + (1-d)\frac{(1-D)}{1-\hat{\pi}_{-k}(z)}$ for each individual, were D denotes the treatment received.

   (c) Solve the following weighted estimating equation;

$$S(\alpha_{dk}, \gamma_{dk}) = \sum_{i \in R_{-k}} W_i^{(d)} \begin{pmatrix} \log\left(\hat{Y}_{-k}^{(d)}(z_i)\right) \\ \tilde{z}_i \end{pmatrix} \left(Y_i - \exp\left(\alpha_{dk} \log\left(\hat{Y}_{-k}^{(d)}(z_i)\right) + \gamma_{dk}^T \tilde{z}_i\right) \times L_i\right) = 0,$$

   where $\tilde{z}$ is z with an intercept; (8)

   by fitting a Poisson regression with Y as the response, log-transformed predictions, $\log\left(Y^{(r)}(z)\right)$ and z as the covariates, the log-transformed offset (log(L)), and $\hat{W}^{(d)}$ from the PS estimation as weights. Denote the solution to the estimating equations by $(\hat{\alpha}_{rk}, \hat{\gamma}_{rk})$.

3. Fit another Poisson regression with $\exp\left(\hat{\alpha}^{rk} \log\left(\hat{Y}_{-k}^{(d)}(z)\right) + \hat{\gamma}_{rk}^T \tilde{z}\right)$ as the response, using only covariates, z and not including an offset or weights. Solve the doubly robust estimating

equation;

$$S(\beta_d) = \sum_{k=1}^{K} \sum_{i \in D_k} \tilde{z}_i \left(\exp\left(\hat{\alpha}_{dk} \log\left(\hat{Y}_{-k}^{(d)}(z_i) \right) + \hat{\gamma}_{dk}^T \tilde{z}_i \right) - \exp(\beta_d^T \tilde{z}_i) \right) = 0, \tag{9}$$

to obtain $\hat{\beta}_d$, representing the treatment effect.

4. Repeat steps (a - c) with B bootstrap samples to obtain $\hat{\beta}_d^b$. The final estimator $\hat{\beta}_d$ is the mean across all bootstrap estimates $\hat{\beta}_d^b$.

5. Compute the Conditional Average Treatment Effect (CATE) for a set of covariates x as $CATE_{tworeg}(z) = (\hat{\beta}_1 - \hat{\beta}_0)^T \tilde{z}$.

Models:

Conventional DGE Analysis

Let $Y_{gi} \overset{i.i.d}{\sim} NB(\lambda_{gi}, \phi_g)$,

Naive Analysis:

$$\log(\lambda_{gi}) = \beta_{0g} + \beta_{1g}\text{SLE\_Status}_i + \log L_i, \ i = 1, 2, \cdots, 256. \tag{10}$$

Conditional Analysis:

$$\log(\lambda_{gi}) = \beta_{0g} + \beta_{1g}\text{SLE\_Status}_i + \beta_{2g}\text{Batch}_i + \beta_{3g}\text{Age}_i + \beta_{4g}\text{Sex}_i + \beta_{5g}\text{PopCov}_i + \log L_i, \tag{11}$$
$$i = 1, 2, \cdots, 256.$$

where;

L_i – Effective Library size of sample i

ϕ_g – Overdispersion parameter for gene, g

$\lambda_{gi} = E(Y_{gi}|L_i, Z_i)$ – mean count given the library size and covariates Z_i in the model.

Standardization Approach

The model in Equation 12 was fitted for the diseased and healthy groups seperately.
Let $Y_{gi} \overset{i.i.d}{\sim} Po(\lambda_{gi})$,

$$\log(\lambda_{gi}) = \beta_{0g} + \beta_{1g}\text{Batch}_i + \beta_{2g}\text{Age}_i + \beta_{3g}\text{Sex}_i + \beta_{4g}\text{PopCov}_i + \log L_i, \tag{12}$$

where $i = 1, 2, \cdots, 158$ in the diseased group and $i = 1, 2, \cdots, 98$ in the healthy group.

IPW Approach

The logistic treatment regression model to compute the weights.

$$\text{logit}(Pr(D_i = 1|Z_i) = \gamma_0 + \gamma_1 \text{Batch}_i + \gamma_2 \text{Age}_i + \gamma_3 \text{Sex}_i + \gamma_4 \text{PopCov}_i + \log L_i, \quad (13)$$
$$i = 1, 2, \cdots, 256.$$

The outcome regression with inverse probabilities as weights.

$$\log(\lambda_{gi}) = \alpha_{0g} + \alpha_{1g} \text{SLE\_status}_i + \log L_i, i = 1, 2, \cdots, 256. \quad (14)$$

Doubly Robust Approach

$$\textbf{cate.model} = \text{Expression} \sim \text{batch\_cov} + \text{Age} + \text{Sex} + \text{PopCov} + \text{offset(LogLibSize)} \quad (15)$$

$$\textbf{ps.model} = \text{Disease} \sim \text{batch\_cov} + \text{Age} + \text{Sex} + \text{PopCov} \quad (16)$$

Results:

The distribution of overdispersion parameter estimates for 100 genes.

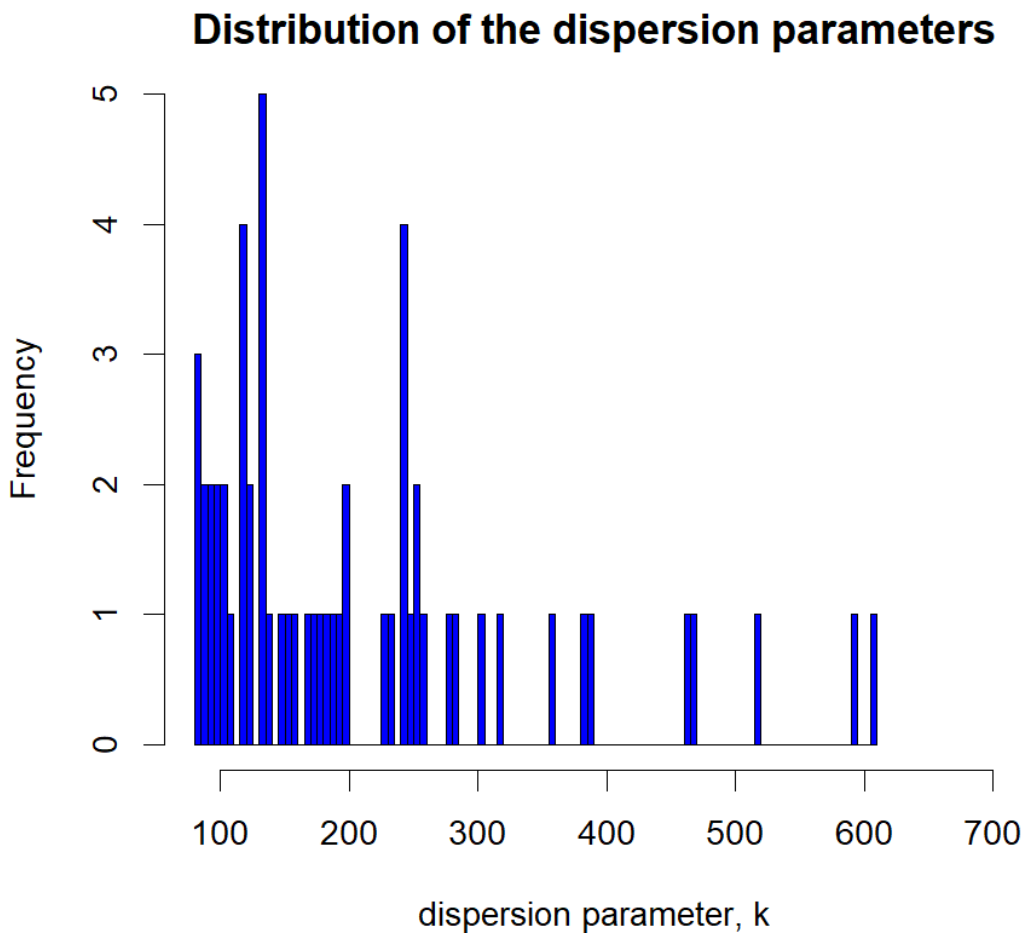


Figure 15: Distribution of the dispersion parameters estimated based on NB-GLM models for 100 genes.

Comparison between the top 100 significant genes based on IPW versus Standardization

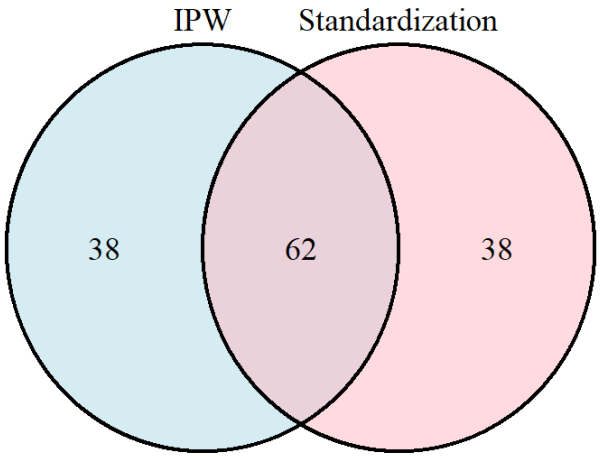


Figure 16: A Venn diagram comparing the top 100 significant genes based on IPW versus Standardization.

Comparison between the top 100 significant genes based on IPW, Standardization and Doubly robust methods.

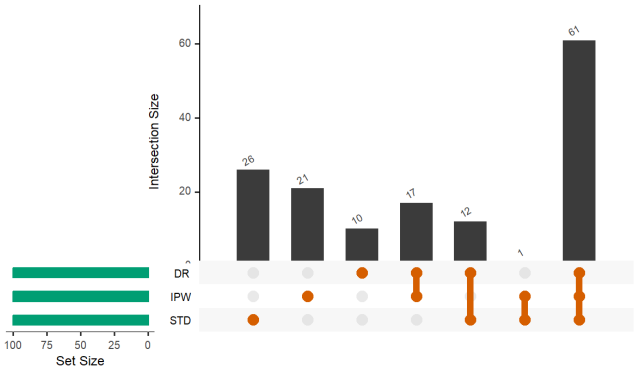


Figure 17: Upset plot comparing the top 100 significant genes based IPW, Standardization and Doubly robust estimation approaches.

R code

```
#********************************** Loading Data ************************************
# Reading Systematic Lupus Erythematosus(SLE) data
sleData <- readRDS("240209_lupusDataSCE_raw_CM.rds")
# Assign named assay "X" to counts: For easy manupulation
counts(sleData) <- assay(sleData, "X")
# dim(counts(sleData)) # 12869 307429

# Aggregation of single-cell to pseudobulk data
#***************************** pseudo-bulk samples***************************************
# Creating pseudo-bulk samples (from cell level counts): Aggregate by ind_cov_batch_cov
sle_summed <- aggregateAcrossCells(sleData,
                                   id=colData(sleData)[,c("ind_cov_batch_cov")])
# dim(counts(sle_summed)) # 12869 genes , 355 samples;
# Number of genes maintained -- The aggregation is done for every gene.

#***********************  Independence assumption!!!  *********************************
# NOTE: For some patients, multiple samples were taken, So no independence......
# -----------selecting one sample per patient from the aggregated data---------#
# --------- Select the samples with the highest total count --------------------#

# Add a column 'total_counts' to the'metadata', representing total counts for each cell
sle_Data_count <- sleData
colData(sle_Data_count)$total_counts <- colSums(counts(sle_Data_count))
metadata_count <- as.data.frame(colData(sle_Data_count))

# Summarize to find the ind_cov_batch_cov with the highest total count
ind_batch_summary <- metadata_count %>%
  group_by(ind_cov, ind_cov_batch_cov) %>%
  summarise(total_counts_sum = sum(total_counts), .groups = 'drop') %>%
  arrange(ind_cov, desc(total_counts_sum)) %>%
  group_by(ind_cov) %>%
  filter(row_number() == 1) %>%
  ungroup()

# Extract the ind_cov_batch_cov identifiers to retain
ind_batch_ids_to_retain <- as.character(ind_batch_summary$ind_cov_batch_cov)

#  Subset the aggregated data to contain one sample per patient
sle_summed_indep <-  sle_summed[, sle_summed$ind_cov_batch_cov %in%
                                  ind_batch_ids_to_retain]

# dim(sle_summed_indep) # 12869 genes,   261 unique samples
```

```
# Delete samples from from African American and Hispanic because there are few
# data points. Only 3 from African American and 2 from Hispanic.
sle_summed_indep <- sle_summed_indep[, colData(sle_summed_indep)$pop_cov %in%
                                       c("Asian", "European")]

# Dropping unused levels in all factor columns in colData
colData(sle_summed_indep) <- DataFrame(lapply(colData(sle_summed_indep), function(x) {
  if (is.factor(x)) droplevels(x) else x
}))
# dim(sle_summed_indep) # 12869   256; NOTE: 5 samples were removed

#******************    Exploration of expression data     *************

# Un Supervised Exploratory Data Analysis
# PCA plot
pcaplot1 <-plotReducedDim(sle_summed_indep, "X_pca", colour_by = "SLE_status")
p1 <- pcaplot1 + labs(x = "PCA 1", y = "PCA 2", title = "PCA Plot by SLE Status") +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))

# UMAP plot
umapplot1 <- plotReducedDim(sle_summed_indep, "X_umap", colour_by = "SLE_status")
p2 <- umapplot1 + labs(x = "UMAP 1", y = "UMAP 2", title = "UMAP Plot of SLE Status") + theme_bw() +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10))
# Combined plot
grid.arrange(p1, p2, nrow = 1, ncol = 2)

#******************** Fitting NB Models using glmmTMB **********************
# Assign the data object 'sle_summed_indep' to the variable 'sle'to work with.
sce <- sle_summed_indep
# FILTERING
# Remove genes that are lowly expressed
keep <- filterByExpr(sce, group=sce$SLE_status) # default min total count 15
# removes genes that are not expressed above a log-CPM threshold in a minimum number of samples
sce <- sce[keep,]; summary(keep) # Remained 8578 genes

# Calculating normalised library sizes (effective library sizes)
sce_norm <- calcNormFactors(sce, method = "TMM") # or use normLibSizes() function which is the same
NormLibSize <- sce_norm$samples$lib.size*sce_norm$samples$norm.factors

# Naive Case: gene expression ~ disease
# Initialize a list to store model results for all genes
models.list <- list()
# Initialize a vector to flag genes with fitting issues
```

```
problematic_genes <- character()
# Create a dataframe for glmmTMB fitting
lupus <- data.frame(
  Expression = NA,
  Disease = sce$SLE_status,
  LogLibSize = log(NormLibSize)
)
# Loop to fit models and catch any fitting problems
for (i in 1: nrow(sce)) {
  gene_name <- rownames(sce)[i]
  # print(paste("Processing gene:", gene_name))
  # Attempt to fit the model, catch warnings and errors
  fit_result <- tryCatch({
    warning_issue <- FALSE
    warning_message <- ""
    # Catch warnings specifically
    model <- withCallingHandlers({
      lupus$Expression <- as.numeric(counts(sce)[i, ])
      # Model
      glmmTMB(Expression ~ Disease + offset(LogLibSize), data = lupus, family = nbinom2(link="log"))
    }, warning = function(w) {
      warning_issue <- TRUE
      warning_message <- w$message
      invokeRestart("muffleWarning")
    })

    list(model = model, issue = warning_issue, message = warning_message)
  }, error = function(e) {
    # Handle errors
    list(model = NULL, issue = TRUE, message = e$message)
  })

  # Store the model result
  models.list[[gene_name]] <- fit_result

  # If there was an issue (warning or error), flag the gene
  if (fit_result$issue) {
    problematic_genes <- c(problematic_genes, gene_name)
    message_type <- ifelse(fit_result$message == "", "Error", "Warning")
    print(paste(message_type, "with gene:", gene_name, "Message:", fit_result$message))
  }
}
# Save the list of full model objects
saveRDS(models.list, "models_list.rds")
```

```r
# Load the models list from the RDS file
mlist <- readRDS("models_list.rds")
# Initialize an empty dataframe for storing Effect size, standard error and p-values
resultsNB_df <- data.frame(
  Gene = character(),
  effect_size = numeric(),
  std_error = numeric(),
  PValue = numeric()
)
# Loop through each model in the list
for (gene in names(mlist)) {
  # Extract the model from the list
  model.fit <- mlist[[gene]]

  # Extract out effect size, standard error and pvalue for each gene
  effect_size <- summary(model.fit$model)$coefficients$cond[2, 1]
  std_error <- summary(model.fit$model)$coefficients$cond[2, 2]
  pvalue  <- summary(model.fit$model)$coefficients$cond[2, 4]
  # Append the gene name, effect size, standard error and its p-value to the dataframe
  resultsNB_df <- rbind(resultsNB_df,
                        data.frame(Gene = gene, EffectSize = effect_size,
                                   StdError = std_error, PValue = pvalue))
}
# Multiplicity correction using BH
resultsNB_df$p.adj <- p.adjust(resultsNB_df$PValue, method = "BH")
sum(resultsNB_df$p.adj > 0.05) # significant genes
# Conditional case: gene expression ~ disease + covariates
# Initialize a list to store model results for all genes
models.list2 <- list()
# Initialize a vector to flag genes with fitting issues
problematic_genes2 <- character()
# Create a dataframe for glmmTMB fitting
lupus <- data.frame(
  Expression = NA,
  Disease = sce$SLE_status,
  batch_cov = factor(sce$batch_cov),
  Age = as.numeric(sce$Age),
  Sex = sce$Sex,
  PopCov = sce$pop_cov,
  LogLibSize = log(NormLibSize)
)
# Loop to fit models and catch any fitting problems
for (i in 1: nrow(sce)) {
  gene_name <- rownames(sce)[i]
  # Attempt to fit the model, catch warnings and errors
```

```
  fit_result <- tryCatch({
    warning_issue <- FALSE
    warning_message <- ""

    # Catch warnings specifically
    model <- withCallingHandlers({
      lupus$Expression <- as.numeric(counts(sce)[i, ])
      # Model
      glmmTMB(Expression ~ Disease + batch_cov + Age + Sex +
                PopCov + offset(LogLibSize), data = lupus, family = nbinom2(link="log"))
    }, warning = function(w) {
      warning_issue <- TRUE
      warning_message <- w$message
      invokeRestart("muffleWarning")
    })
    list(model = model, issue = warning_issue, message = warning_message)
  }, error = function(e) {
    # Handle errors
    list(model = NULL, issue = TRUE, message = e$message)
  })
  # Store the model result
  models.list2[[gene_name]] <- fit_result
  # If there was an issue (warning or error), flag the gene
  if (fit_result$issue) {
    problematic_genes <- c(problematic_genes, gene_name)
    message_type <- ifelse(fit_result$message == "", "Error", "Warning")
    print(paste(message_type, "with gene:", gene_name, "Message:", fit_result$message))
  }
}
# Save the list of full model objects
saveRDS(models.list2, "models_list2.rds")

# Load the models list from the RDS file
mlist_all <- readRDS("models_list2.rds")
# Initialize an empty dataframe for storing Effect size, standard error and p-values
resultsNB_df2 <- data.frame(
  Gene = character(),
  effect_size = numeric(),
  std_error = numeric(),
  PValue = numeric()
)
# Loop through each model in the list
for (gene in names(mlist_all)) {
  # Extract the model from the list
  model.fit <- mlist_all[[gene]]
```

```
  # Extract out effect size, standard error and pvalue for each gene
  effect_size <- summary(model.fit$model)$coefficients$cond[2, 1]
  std_error <- summary(model.fit$model)$coefficients$cond[2, 2]
  pvalue  <- summary(model.fit$model)$coefficients$cond[2, 4]
  # Append the gene name, effect size, standard error and its p-value to the dataframe
  resultsNB_df2 <- rbind(resultsNB_df2,
                         data.frame(Gene = gene, EffectSize = effect_size,
                                    StdError = std_error, PValue = pvalue))
}
# Multiplicity correction using BH
resultsNB_df2$p.adj<-p.adjust(resultsNB_df2$p.adj, method = "BH")
sum(resultsNB_df2$p.adj < 0.05, na.rm = TRUE) # significant genes
#********************************** END-CODE ****************************************
```

The additional R code for the analysis has been uploaded to GitHub i.e, click Code_Anthony_Thesis.